VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY

FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

IMPLEMENTACE A APLIKACE STATISTICKÝCH METOD VE VÝZKUMU, VÝROBNÍ TECHNOLOGII A ŘÍZENÍ JAKOSTI

IMPLEMENTATION AND APPLICATION OF STATISTICAL METHODS IN RESEARCH, MANUFACTURING TECHNOLOGY AND QUALITY CONTROL

DIZERTAČNÍ PRÁCE
DOCTORAL THESIS

AUTOR PRÁCE — Ing. KAREL KUPKA
AUTHOR

VEDOUCÍ PRÁCE — doc. RNDr. ZDENĚK KARPÍŠEK, CSc.
SUPERVISOR

BRNO 2011

Zadání disertační práce

Vytvoření a analýza statistických postupů pro účely a současné potřeby analýzy dat v technologických aplikacích a v oblasti řízení kvality, jejich implementace a použití v reálných studiích na reálných datech.

Abstrakt

Práce se zabývá možnostmi použití moderních statistických postupů se zaměřením na robustní metody. Vybrané postupy jsou analyzovány a aplikovány na častých problémech z praxe v českém průmyslu a technologii. Studovaná témata, metody a algoritmy jsou voleny tak, aby byla přínosem v reálných aplikacích ve srovnání s používanými klasickými metodami. Použitelnost a účinnost algoritmů je ověřena a demonstrována na reálných studiích a problémech z výzkumného prostředí českých průmyslových subjektů. V práci je poukázáno na nevyužitý potenciál současné teoreticko-matematické a výpočetní kapacity a nových přístupů k chápání statistických modelů a metod. Výsledkem práce je rovněž původní vývojové prostředí s programovacím jazykem DARWin (Data Analysis Robot for Windows) pro intenzivní využití efektivních numerických postupů pro získávání informací z dat. Práce je impulsem pro širší využití robustních a numericky, nebo výpočetně náročnějších metod, jako jsou neuronové sítě, pro modelování procesů a kontrolu kvality.

Abstract

This thesis deals with modern statistical approaches and their application aimed at robust methods and neural network modelling. Selected methods are analyzed and applied on frequent practical problems in czech industry and technology. Topics and methods are to be benificial in real applications compared to currently used classical methods. Applicability and effectivity of the algorithms is verified and demonstrated on real studies and problems in czech industrial and research bodies. The great and unexploited potential of modern theoretical and computational capacity and the potential of new approaces to statistical modelling and methods. A significant result of this thesis is also an environment for software application development for data analysis with own programming language DARWin (Data Analysis Robot for Windows) for implemenation of effective numerical algorithms for extaction information from data. The thesis should be an incentive for boarder use of robust and computationally intensive methods as neural networks for modelling processes, quality control and generally better understanding of nature.

Klíčová slova:

Řízení jakosti, směs rozdělení, robustní metody, robustní regrese, regrese se zlomem, M-odhady, Lp-odhady, bod změny, neuronové sítě

Keywords:

Quality control, normal mixture, robust methods, robust regression, segmented regression, M-estimates, L_p-estimates, change point, neural networks

Citace práce:

KUPKA, K. Implementace a aplikace statistických metod ve výzkumu, výrobní technologii a řízení jakosti. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2011. XY s. Vedoucí dizertační práce doc. RNDr. Zdeněk Karpíšek, CSc..

Prohlášení autora:

Prohlašuji, že jsem disertační práci zpracoval samostatně a použil jsem pouze literaturu, která je uvedena v bibliografii.

Podpis autora:

Poděkování:

Za cenné rady a připomínky k tématu děkuji svému vedoucímu doc. RNDr. Zdeňku Karpíškovi CSc., VUT Brno, prof. Robertovi Vanderbei, Princeton University, prof. Danielovi Jeskemu, University of California Riverside, prof. Jaromíru Antochovi, Univerzita Karlova, Praha, Prof. Andrew R. Barronovi, Yale University, prof. Billu Venablesovi, CSIRO Australia.

1. Obsah

2. Úvod

V posledních letech neustále roste potřeba statistické analýzy experimentálních dat v technice, technologii, výrobě a výzkumu současně s rostoucím obecným povědomím a zájmem o tuto oblast. Matematická statistika a statistická analýza je obor výrazně aplikační a má dlouhou historii. Významným aplikačním polem statistických metod jsou technologie, aplikovaný výzkum a hodnocení kvality výroby. Za počátky systematického využití statistických metod ve výzkumu technologii bývají uváděna dvacátá a třicátá léta dvacátého století v souvislosti s autory jako Ronald A. Fisher [1], John W. Tukey [3] [204], George Box, Walter A. Shewhart [4] za výrazné popularizace a spoluvytváření metodik zavádění principů statistického myšlení („statistical thinking“) do (především technologické) praxe autory, jako byli například Joseph Juran, W. E. Deming[6]. Všeobecně přijaté principy hodnocení kvality procesů pomocí konceptu ztrátové funkce $Loss(X) \sim (X - T)^2$ jako čtverce vzdálenosti od požadované hodnoty T logicky ospravedlnilo široké nasazení statistických metod [4] - [16], v nichž je ústředním tématem rozptyl, jeho rozklad, vysvětlení a minimalizace. Jedna z celosvětově nejpoužívanějších statistik pro vyjádření kvality je například převrácená hodnota výběrové směrodatné odchylky $c_p = k\hat{\sigma}^{-1}$ nazývaná index způsobilosti. Každé oprávněné vysvětlení variability je tak žádoucí a lze je považovat za potenciální zvýšení kvality a tím zisku, konkurenceschopnosti, atd., neboť vysvětlení variability znamená netriviální matematický model, který popisuje nový fyzikální mechanismus, přináší nové poznatky a někdy reálně vede k pochopení, ovlivnění či zpřesnění experimentu, procesu, apod. K účelu vysvětlování variability se velmi dobře hodí matematicko-statistické modely a postupy, obecně především v širším pojetí regrese a klasifikace. Důležitým aspektem je přitom důraz na reálné podmínky průmyslových technologií a z nich plynoucí data, která často nevyhovují předpokladům kladeným na výběry při klasických statistických metodách. Od nejjednodušších základů v podobě Shewhartových regulačních diagramů a jejich mnoha modifikací se prosazují další nástroje, především analýza rozptylu, optimalizace experimentu, regrese, časové řady, vícerozměrné metody a různé exploratorní a prediktivní techniky (Data Mining, neuronové sítě, Support Vector Machines). Stále podceňovaným aspektem statistického vyhodnocování pozorování jsou předpoklady, za kterých jsou statistické metody použitelné. Reálná data často zdaleka neodpovídají normálnímu rozdělení, nejsou nezávislá, homoskedastická, homogenní, lineární [2]. Místo snahy o pochopení fyzikálních mechanismů a modelů, korektní a smysluplné vysvětlení pozorovaného fenoménu, se v praxi setkáváme s mechanickým používáním několika naučených vztahů jako nutného administrativního úkonu vedoucímu k misinterpretaci dat, chybným a nepoužitelným výsledkům a smluvním a právním konfliktům. Místo hlubšího pochopení studovaných problémů vede nedostatečné statistické vzdělání absolventů technických a přírodovědných oborů a používání nevhodných postupů nebo statistik k nedůvěře, až pohrdání statistikou jako matematickým oborem, který má ovšem více než jiné oblasti aplikované matematiky potenciál propojit nejen vědecké, a tedy experimentální obory s matematikou, ale i propojit zdánlivě zcela nesouvisející vědecké

obory mezi sebou a umožnit tak masivní mezioborovou komunikaci a výměnu informací a zkušeností na bázi podobných statistických modelů a metod. Jak píše John Tukey již v roce 1949 [180]: „Statistika jako doktrína pro plánování experimentů a pozorování a pro interpretování dat má společný vztah ke všem vědám.“

Tato práce se zabývá popisem několika statistických metod a postupů a uvádí výsledky jejich aplikace v technologii a aplikovaném výzkumu v podnicích a institucích v České republice. Pro účely a potřeby implementace statistických metod a jejich reálné nasazení byl autorem vyvinut statistický systém QCExpert a v jeho rámci i programovací jazyk DARWin (Data Analysis Robot for Windows). Tento jazyk je používán i v této práci pro kodifikaci a dokumentaci používaných algoritmů, a proto je jeho popis uveden na konci práce jako příloha. Všechny výpočty, algoritmy a grafy uvedené v této práci jsou vytvořeny v rámci tohoto systému a systém QCExpert a DARWin je autorův příspěvek k širšímu používání statistických metod v praxi a výuce.

3. Jednorozměrné rozdělení

Jedním z prvních problémů při popisu technologických dat jsou odhady parametrů předpokládaného jednorozměrného rozdělení $F(x)$, jsou-li data zatížena někdy obtížně identifikovatelnými vybočujícími hodnotami, či obecněji hodnotami pocházejících z jiných fyzikálních procesů, a tedy zřejmě i z jiného rozdělení. Problém identifikace takových hodnot nemusí být vždy triviální, neboť mohou ležet i blízko střední hodnoty $F(x)$ a nemusí být na první pohled patrné. Takové případy jsou zmíněny dále. Postupy analýzy takovýchto heterogenních dat zahrnují testy a filtry na odlehlé hodnoty, použití robustních metod, nebo modely obsahující více rozdělení. Dva poslední přísupy budou zmíněny a použity v následujících odstavcích. Z robustních metod se budeme věnovat především Lp-odhadům, a M-odhadů, které jsou předmětem výzkumů a aplikací v řadě oborů, např. [75] - [81] a [184] - [205].

3.1. Lp odhady

3.1.1. Úvod

Vycházíme z náhodné proměnné X pocházející z normálního rozdělení definovaného hustotou

$$f\left(\mathbf{x};\mu\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right] \qquad \textbf{(1)}$$

Budeme-li hledat maximálně věrohodný odhad střední hodnoty za předpokladu konstantního rozptylu, dostaneme

$$\ln\underset{\max(\mu)}{L}\left(\mathbf{x};\mu\right)=-n\ln\left(\sigma\sqrt{2\pi}\right)-\frac{n}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2.$$

Po vynechání konstant, které nemají vliv na polohu maxima a obrácení znamémka máme jednoduchou úlohu minimalizace čtverců odchylek (metoda nejmenších čtverců, L_2 - norma), která je řešitelná analyticky

$$-\ln L(\mathbf{x};\mu) \underset{\min(\mu)}{=} \sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}\left(x_i^2 - 2\mu x_i + \mu^2\right)$$

Položíme-li derivaci rovnu nule, získáme podmínku minima

$$\frac{\partial \ln L}{\partial \mu} = 2\sum_{i=1}^{n}(\mu - x_i) = n\mu - \sum_{i=1}^{n} x_i = 0 \qquad \textbf{(2)}$$

a odtud maximálně věrohodný odhad

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textbf{(3)}$$

Připomeňme, že změnou exponentu r členu $|x - \mu|^r$ ve vztahu pro hustotu můžeme získat dvě další v praxi velmi často používaná limitní rozdělení – Laplaceovo pro $r = 1$ a rovnoměrné pro $r \to \infty$. Podobným postupem pak získáme pro Laplaceovo rozdělení

$$f(\mathbf{x};\mu) = \frac{1}{2\lambda}\exp\left[-\frac{|x-\mu|}{\lambda}\right] \qquad \textbf{(4)}$$

maximálně věrohodný odhad μ minimalizací absolutních odchylek (L_1 normy)

$$-\ln L(\mathbf{x};\mu) \underset{\min(\mu)}{=} n\ln 2\lambda + \frac{1}{\lambda}\sum_{i=1}^{n}|x_i-\mu| \qquad \textbf{(5)}$$

Setřídí-li se hodnoty x_i, pak lze bez újmy na obecnosti rozdělit data na dvě skupiny

$$\left\{\begin{array}{c} x_{(i)} < \mu \ pro\ i = 1,\ldots,k \\ x_{(i)} > \mu \ pro\ i = k+1,\ldots,n \end{array}\right\}$$

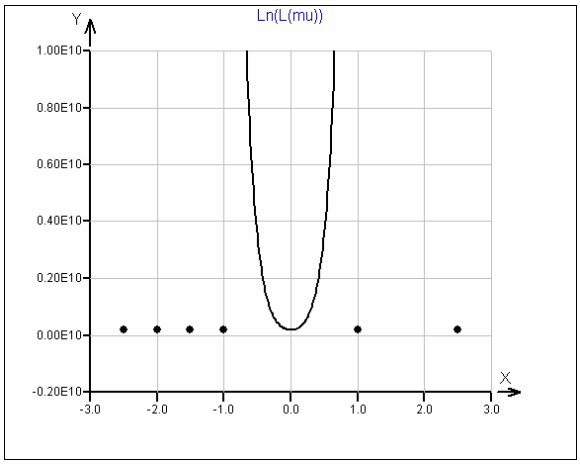
a minimalizovat

$$\sum_{i=1}^{k}\mu - x_{(i)} + \sum_{i=k+1}^{n} x_{(i)} - \mu = (2k-n)\mu - \sum_{i=1}^{k} x_{(i)} + \sum_{i=k+1}^{n} x_{(i)}\ .$$

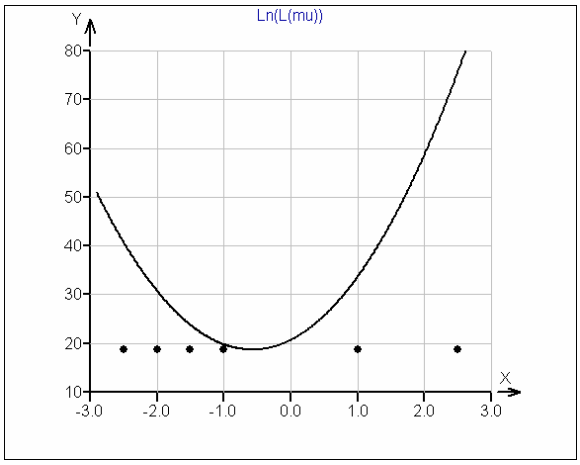
položením derivace $(2k-n)$ rovné nule se získá maximálně věrohodný odhad μ v podobě podmínky $k = \frac{n}{2}$, kterou splňuje medián. Podobně pro r → ∞ minimalizujeme

$\lim_{r\to\infty}\sum_{i=1}^{n}\left|x_{(i)}-\mu\right|^r$. Označme $a = \min(x_i)$, $b = \max(x_i)$, pak protože

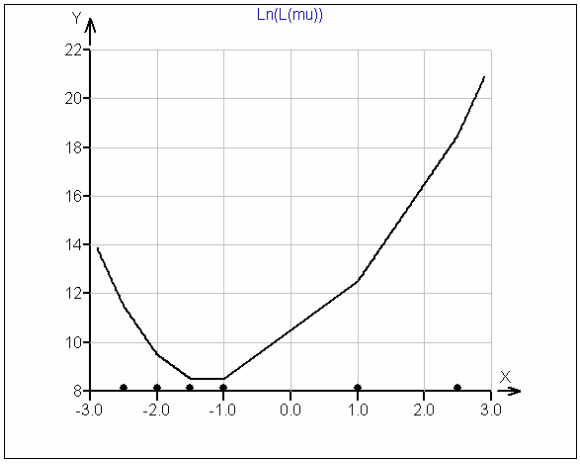
$\lim_{r\to\infty}\frac{\left|x_{(i)}-\mu\right|^r}{\left|a-\mu\right|^r}=0$ *pro* $x_{(i)} > a$ a $\lim_{r\to\infty}\frac{\left|x_{(i)}-\mu\right|^r}{\left|b-\mu\right|^r}=0$ *pro* $x_{(i)} < b$, stačí minimalizovat maximální odchylku (L∞-norma) $\left|a-\mu\right|^r+\left|b-\mu\right|^r$, tedy $\left|a-\mu\right|=\left|b-\mu\right|$ a protože $\left|a-\mu\right|=\mu-a$ a $\left|b-\mu\right|=b-\mu$ plyne odtud $\mu=\frac{b+a}{2}$, což je vztah pro polosumu. Obecně existuje pro celá r > 2 celkem (r – 1) řešení, z nichž však naštěstí jen jedno je reálné, ostatní mají nenulovou imaginární složku a nelze je interpretovat jako střední hodnotu reálné proměnné.



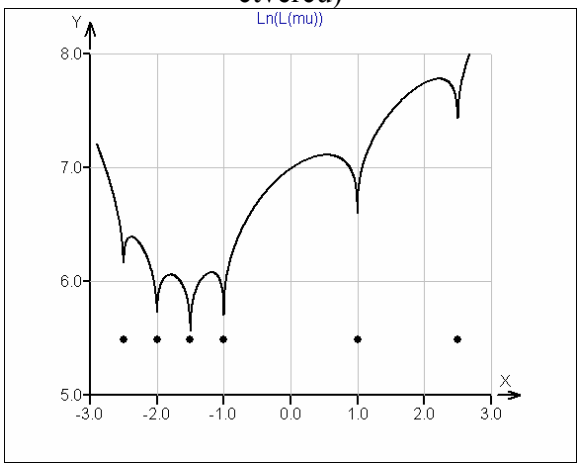
A: ln$L(\mu)$, $r = 20$



B: ln$L(\mu)$, $r = 2$ (metoda nejmenších čtverců)



C: ln$L(\mu)$, $r = 1$ (metoda nejmenších absolutních odchylek)



D: ln$L(\mu)$, $r = 0.2$

Obr. 1 Minimum věrohodnosti pro Lp-odhady

Laplaceovo rozdělení s těžšími konci než normální rozdělení se používá jako model pro výběry, které sice mohou pocházet z normálního rozdělení, v nichž se však vyskytují odlehlé hodnoty pocházející z jiných rozdělení, obvykle s různými rozptyly, ale stejnou, nebo nepříliš odlišnou střední hodnotou. Medián je jako odhad střední hodnoty vůči

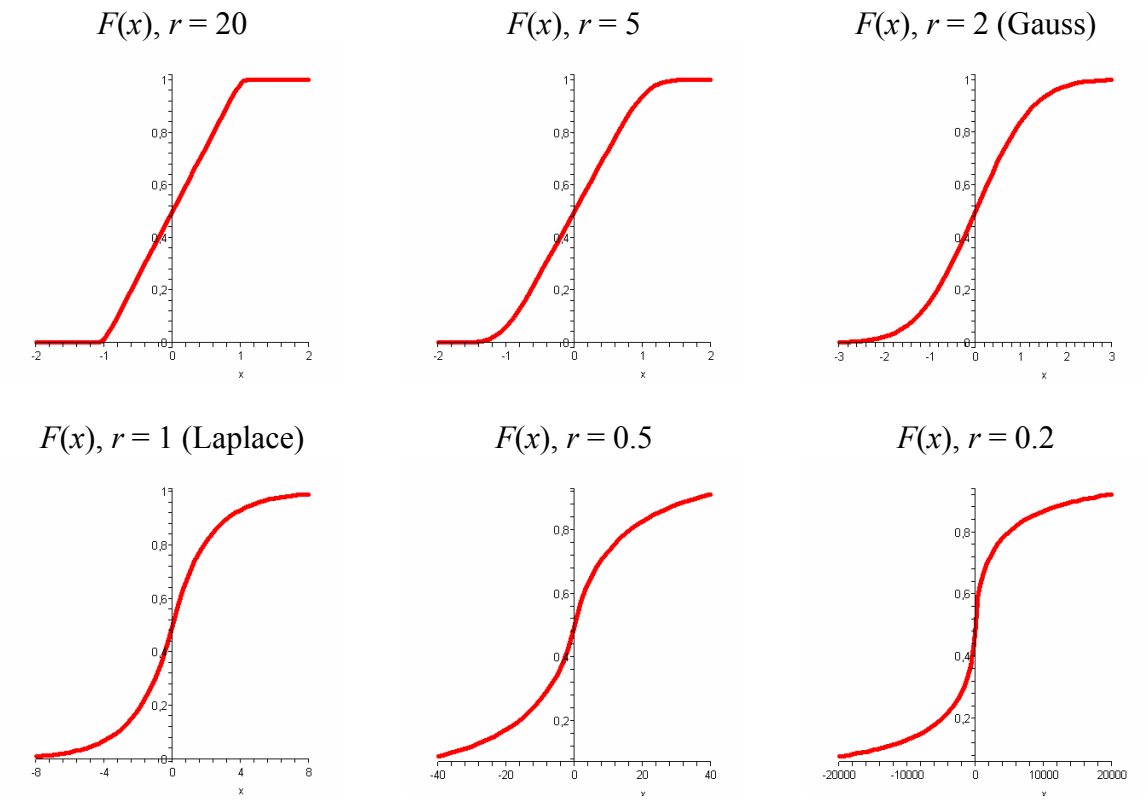
takovým odlehlým měřením robustní. Odlišné chování má polosuma, která se používá například při odhadu střední hodnoty u dat s malou náhodnou chybou ale s velkými zaokrouhlovacími chybami, s čímž se lze často setkat při odečítání z displeje s malým počtem desetinných míst, kdy mají takto uměle vzniklé odchylky rovnoměrné rozdělení. Grafy věrohodnosti $\ln L(\mu)$ pro exponenty $r = 20$, $r = 2$ a $r = 1$ jsou znázorněny na Obr. 1 (A – C) pro výběr **x**: (–2.5, –2, –1.5, –1, 1, 2.5).

Při výpočtu mediánu je někdy překážkou ploché minimum, tedy nejednoznačný bodový odhad při sudém n. Hustotu pravděpodobnosti uvedené třídy rozdělení lze obecně zapsat jako

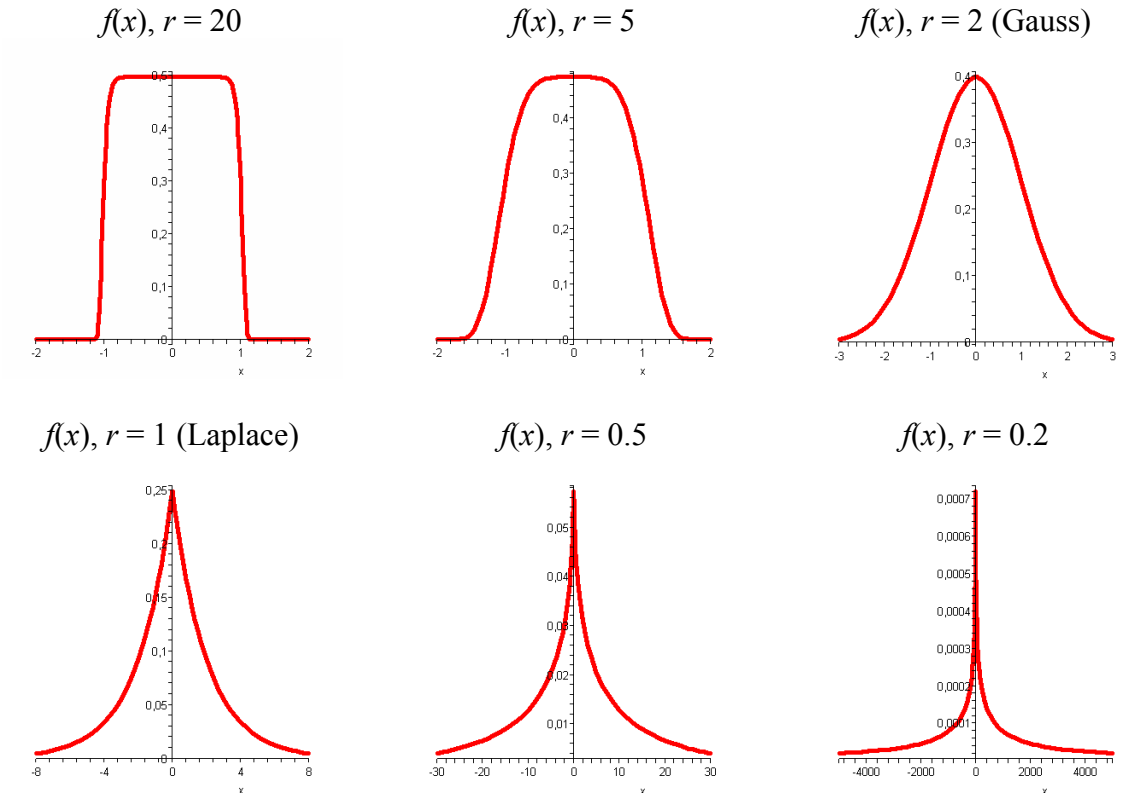
$$f(x) = \frac{r}{2^{\frac{r+1}{r}} L^{\frac{1}{r}} \Gamma\left(\frac{1}{r}\right)} \exp\left(-\frac{|x-\mu|^r}{2L}\right). \qquad (6)$$

kde pro $r = 2$ dostáváme normální, pro $r = 1$ Laplaceovo a pro $r \to \infty$ rovnoměrné rozdělení. Při hodnotě $0 < r < 1$, má $f(x)$ stále konečný (jednotkový) integrál a dokonce i konečné momenty

$$\mu_k = \int_{-\infty}^{\infty} x^k f(x)\,\mathrm{d}x = \frac{1+(-1)^k}{2} \frac{2^{\frac{k-r}{r}} L^{\frac{k}{r}} \Gamma\left(\frac{k+1}{r}\right)}{\Gamma\left(\frac{1}{r}\right)}. \qquad (7)$$



Obr. 2 Distribuční funkce pro různé hodnoty exponentu *r*



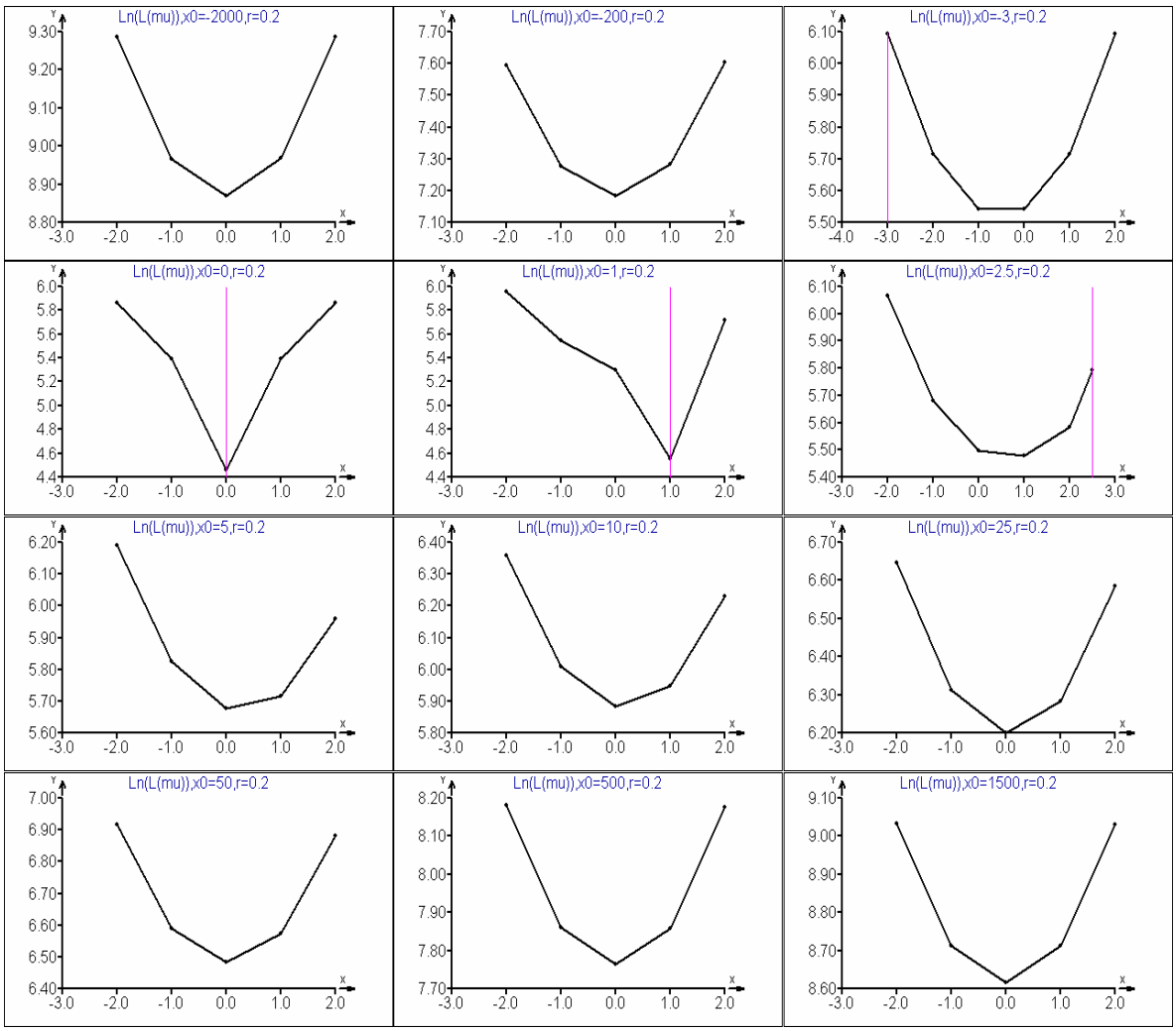
Obr. 3 Hustota pravděpodobnosti pro různé hodnoty exponentu *r*

Věrohodnost $\ln L(\mu)$ je pro $r < 1$ sice po částech striktně konkávní a v minimech nemá derivaci, avšak k nalezení $\min \ln L(\mu)$ je možné využít skutečnosti, že všechna existující lokální minima leží v x_i. Stačí tedy vyšetřit $\ln L(\mu)$ v bodech $x = x_i$ a vybrat nejmenší hodnotu $\min_i \ln L(x_i)$, viz Obr. 1 (D), kde je zobrazena funkce $\ln L(\mu)$ pro $r = 1/5$. Takto získané odhady mohou být robustnější, než medián, jak je ilustrováno v Tab. 1 na příkladu s daty $(-2, -1, 0, 1, 2, x_0)$, kde x_0 je vybočující hodnota (outlier), která nabývá hodnoty od -2000 do +1500. Správná střední hodnota by tedy měla být $\mu = 0$. Ve sloupcích „metody odhadu“ jsou porovnány nerobustní ($r \geq 2$) a robustní ($r < 2$) odhady. Medián je polohou bodu x_0 ovlivněn, kdežto odhad při $r < 1$ zůstává při $x_0 << \mu$ a při $x_0 >> \mu$ nezměněn. Na Obr. 4 je ilustrována změna tvaru věrohodnostní funkce při hodnotách x_0 blízkých střední hodnotě výběru. Někteří autoři [75], [78] však upozorňují na to, že optimum je třeba hledat spíše v nejnižším lokálním maximu $-\ln L(x)$, kde je možné využít standardní derivační minimalizační postupy s kritériem $+\ln L(x)$. Obecně jsou metody L_p pro $p < 1$ málo prozkoumané a publikované.

Popsaný postup je možné použít v podobných situacích jako medián, pokud očekáváme v datovém souboru výskyt velmi odlehlých hodnot, a z nějakých důvodů není možné, nebo účelné odlehlé hodnoty identifikovat, nebo je výběrové rozdělení výrazně leptokurtické (vyšší špičatost, než 3, resp. 6). V následujícím odstavci se zmíníme o možné jiné příčině odchylek od normálního rozdělení.

Tab. 1 Porovnání robustnosti odhadů

	Metody odhadu μ				
Poloha outlieru x_0	polosuma, $r \to \infty$	průměr, r=2	medián, r=1	odhad pro r=0.5	odhad pro r=0.1
-2000	-999	-333.3333	-0.5	0	0
-200	-99	-33.3333	-0.5	0	0
-3	-0.5	-0.5	-0.5	-0.5	-0.5
0	0	0	0	0	0
1	0	0.1667	0.5	1	1
2.5	0.25	0.4167	0.5	1	1
5.0	1.5	0.8333	0.5	0	0
10.0	4	1.6667	0.5	0	0
25.0	11.5	4.1667	0.5	0	0
50.0	24	8.3333	0.5	0	0
500.0	249	83.3333	0.5	0	0
1500.0	749	250.0	0.5	0	0



Obr. 4 Deformace ln$L(\mu)$ při „průletu“ outlieru pro r = 0.2. Poloha outlieru x_0 je znázorněna svislou přímkou uvedena v záhlaví grafů. Hodnoty ln$L(\mu)$ jsou pro názornost spojeny úsečkami

3.1.2. Aplikace

Světový výrobce ocelových a hliníkových automobilových kol Hayes-Lemmerz používá k průběžnému sledování stability produkce geometrické vzdálenosti definovaných bodů na výrobku pomocí prostorového měřidla s přesností 0.001mm. Vybrané vzdálenosti jsou používány ke konstrukci regulačních diagramů s mezemi odvozenými od směrodatné odchylky za předpokladu normality, homogenity, konstantního rozptylu a střední hodnoty. Vybrané části dat jsou uvedeny v Tab. 2. Reálně však data obsahují mírně vybočující data způsobená nečistotami, nebo chybnou kalibrací a silně vybočující data způsobená chybami obsluhy. Klasické odhady zde systematicky selhávají. Bylo proto využito pro odhad střední hodnoty mediánu a pro odhad směrodatné odchylky mediánové směrodatné odchylky $\hat{\sigma}_{MAD}$, pro níž platí

$$\hat{\sigma}_{MAD} = \frac{MAD}{0.6745} = \frac{\text{median}\left|x - \text{median}\left(x\right)\right|}{0.6745}, \qquad \textbf{(8)}$$

kde *MAD* je medián absolutních odchylek od mediánu a $F^{-1}(3/4) = 0.6744897\ldots$ je konstanta, která zajišťuje nestrannost tohoto odhadu směrodatné odchylky. Meze regulačních diagramů na Obr. 5B až Obr. 8B jsou určeny jako median(x) ± $3\,\hat{\sigma}_{MAD}$. V grafech na Obr. 5A až Obr. 8A jsou regulační meze stanoveny klasicky pomocí průměru a směrodatné odchylky $\bar{x} \pm 3\hat{\sigma}, \hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Porovnáním grafů A a B na Obr. 5 až Obr. 8 je zřejmé, že meze založené na i jednoduchých a v praxi snadno aplikovatelných robustních statistikách mají výrazně lepší schopnost identifikovat neplatná data.

Tab. 2 Data z kontroly geometrické stability

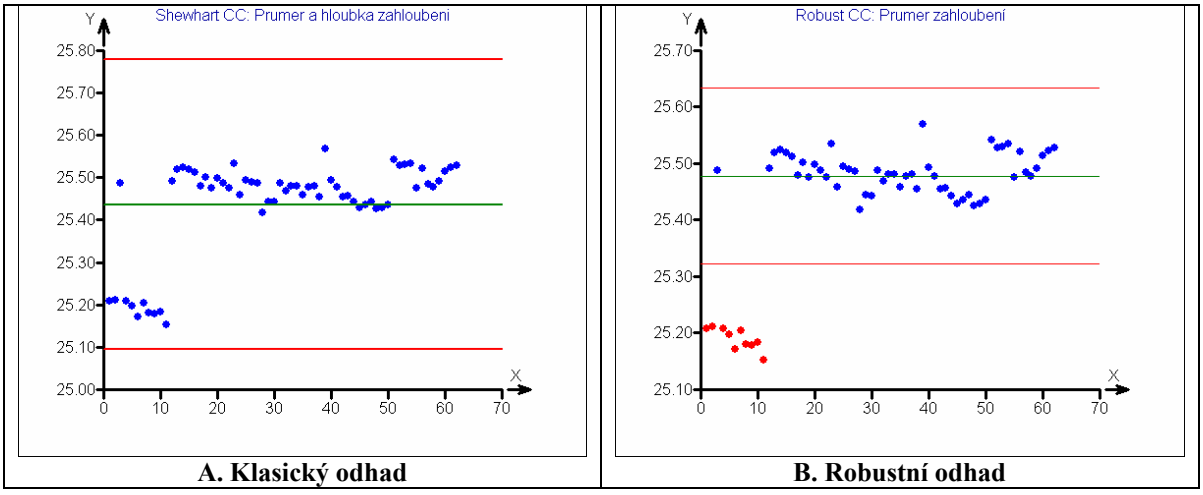
Prumer a hloubka zahloubeni (62 dat):
25.208, 25.211, 25.487, 25.208, 25.197, 25.172, 25.205, 25.18, 25.179, 25.184, 25.152, 25.492, 25.519, 25.524, 25.519, 25.513, 25.48, 25.501, 25.476, 25.498, 25.487, 25.476, 25.534, 25.459, 25.495, 25.489, 25.486, 25.418, 25.444, 25.442, 25.487, 25.469, 25.481, 25.481, 25.458, 25.477, 25.481, 25.454, 25.569, 25.493, 25.478, 25.455, 25.456, 25.443, 25.428, 25.436, 25.444, 25.426, 25.428, 25.436, 25.542, 25.528, 25.53, 25.534, 25.476, 25.521, 25.484, 25.478, 25.492, 25.514, 25.523, 25.528

Krytka Vzdálenost XY Tloustka (90 dat):
3.148, 3.087, 3.163, 3.084, 3.152, 3.081, 3.093, 3.08, 3.099, 3.161, 3.132, 3.126, 3.158, 3.098, 3.148, 3.104, 3.156, 3.107, 3.154, 3.175, 3.116, 3.106, 3.159, 3.133, 3.108, 3.128, 3.137, 3.106, 3.086, 3.099, 3.138, 3.102, 3.146, 3.068, 3.072, 3.146, 3.136, 3.113, 3.128, 3.153, 3.118, 3.108, 3.147, 3.136, 3.116, 3.121, 3.152, 3.152, 3.132, 3.13, 3.107, 3.128, 3.114, 3.111, 3.118, 3.119, 3.108, 3.114, 3.145, 3.141, 3.127, 3.125, 3.098, 3.147, 3.16, 3.129, 3.118, 3.145, 3.136, 3.13, 3.12, 3.142, 3.147, 3.129, 3.128, 3.102, 3.152, 3.161, 3.135, 3.126, 3.106, 3.141, 3.078, 3.067, 3.105, 3.118, 3.146, 3.127, 3.13, 3.13

Prumer odlehceni naboje (147 dat):
60.821, 60.942, 60.869, 60.952, 60.872, 60.945, 60.852, 60.945, 60.869, 60.863, 60.856, 60.615, 60.623, 60.859, 60.782, 60.612, 60.783, 60.778, 60.61, 60.59, 60.806, 60.907, 60.86, 60.855, 60.851, 60.845, 60.948, 60.802, 60.579, 60.807, 60.619, 60.568, 60.806, 60.849, 60.81, 60.8, 60.811, 60.95, 60.948, 60.835, 60.842, 60.843, 60.796, 60.809, 60.852, 60.813, 60.801, 60.849, 60.952, 60.858, 60.869, 60.821, 60.861, 60.607, 60.603, 60.585, 60.865, 60.854, 60.871, 60.867, 60.866, 60.875, 60.855, 60.862, 60.864, 60.858, 60.853, 60.776, 60.815, 60.824, 60.868, 60.818, 60.82, 60.855, 60.937, 60.878, 60.86, 60.961, 60.96, 60.859, 60.941,

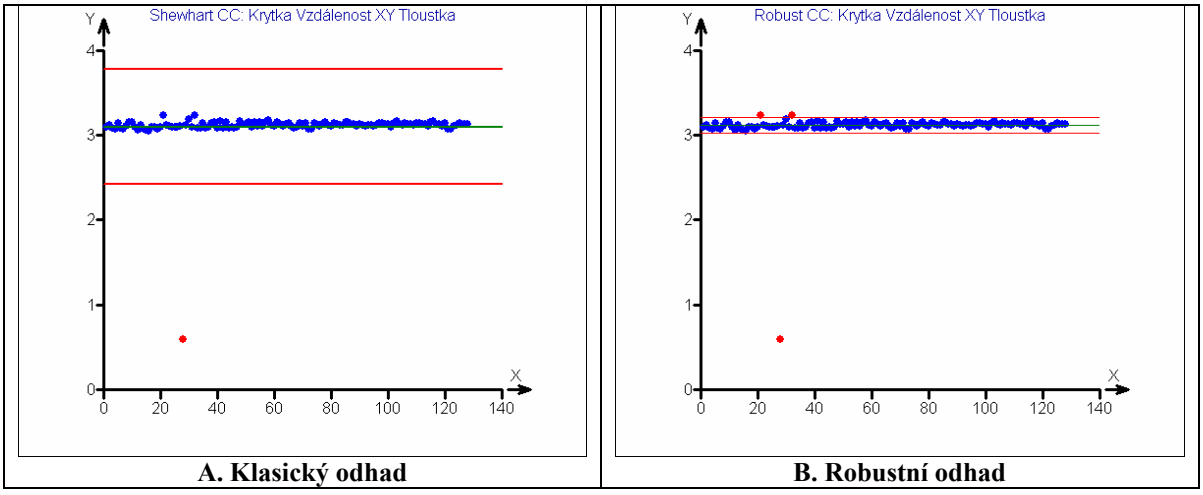
60.86, 60.861, 60.854, 60.862, 60.884, 60.851, 60.954, 60.861, 60.858, 60.946, 60.858, 60.948, 60.856, 60.864, 60.871, 60.944, 60.851, 60.842, 60.87, 60.948, 60.871, 60.808, 60.858, 60.946, 60.862, 60.864, 60.855, 60.939, 60.849, 60.946, 60.595, 60.601, 60.93, 60.935, 60.86, 60.86, 60.938, 60.868, 60.857, 60.944, 60.794, 60.87, 60.748, 60.865, 60.944, 60.759, 60.856, 60.767, 60.86, 60.781, 60.765, 60.858, 60.763, 60.871, 60.884, 60.951, 60.811, 60.85, 60.582, 60.864, 60.935, 60.869, 60.951, 60.863, 60.934, 60.859,

Vzdálenost hrany naboje (151 dat):
12.284, 12.697, 12.68, 12.698, 12.246, 12.164, 12.164, 12.28, 13.112, 12.255, 12.264, 12.238, 12.26, 12.234, 12.191, 12.66, 12.505, 12.285, 12.185, 12.271, 12.275, 12.227, 12.05, 12.21, 12.217, 12.038, 12.053, 12.233, 12.042, 12.247, 12.216, 12.231, 12.22, 12.231, 12.215, 12.051, 12.227, 12.238, 12.051, 12.256, 12.029, 12.265, 12.25, 12.251, 12.056, 12.276, 12.255, 12.243, 12.031, 12.269, 12.284, 12.297, 12.034, 12.298, 12.286, 12.28, 12.056, 12.289, 12.047, 12.692, 12.716, 12.041, 12.056, 12.274, 12.243, 12.064, 12.27, 12.256, 12.077, 12.246, 12.254, 12.341, 12.234, 12.16, 12.366, 12.263, 12.271, 12.24, 12.284, 12.3, 12.251, 12.298, 12.258, 12.224, 12.046, 12.24, 12.278, 12.674, 12.255, 12.058, 12.227, 12.042, 12.257, 12.038, 12.222, 12.018, 12.24, 12.034, 12.274, 12.067, 12.289, 12.291, 12.05, 12.31, 12.307, 12.245, 12.277, 12.247, 12.657, 12.663, 12.261, 12.273, 12.056, 12.269, 12.068, 12.272, 12.063, 12.054, 12.239, 12.246, 12.058, 12.669, 12.291, 12.281, 12.294, 12.273, 12.28, 12.052, 12.059, 12.266, 12.652, 12.311, 12.276, 12.659, 12.28, 12.651, 12.65, 12.295, 12.44, 12.256, 12.272, 12.392, 12.725, 12.282, 12.287, 12.419, 12.439, 12.299, 12.436, 12.296, 12.438



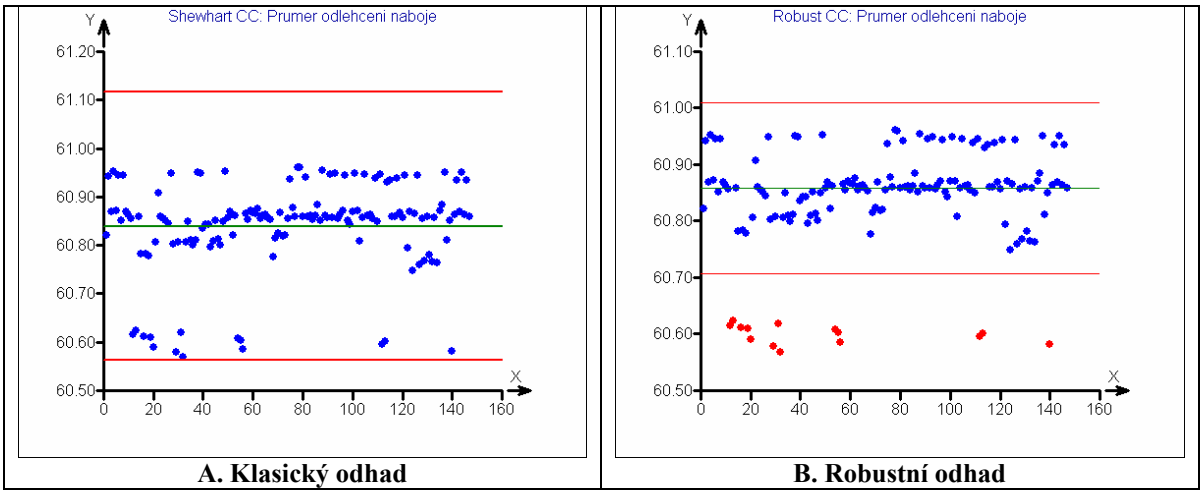
A. Klasický odhad **B. Robustní odhad**

Obr. 5 Odhad střední hodnoty a regulačních mezí klasickým a robustním postupem pro zahloubení



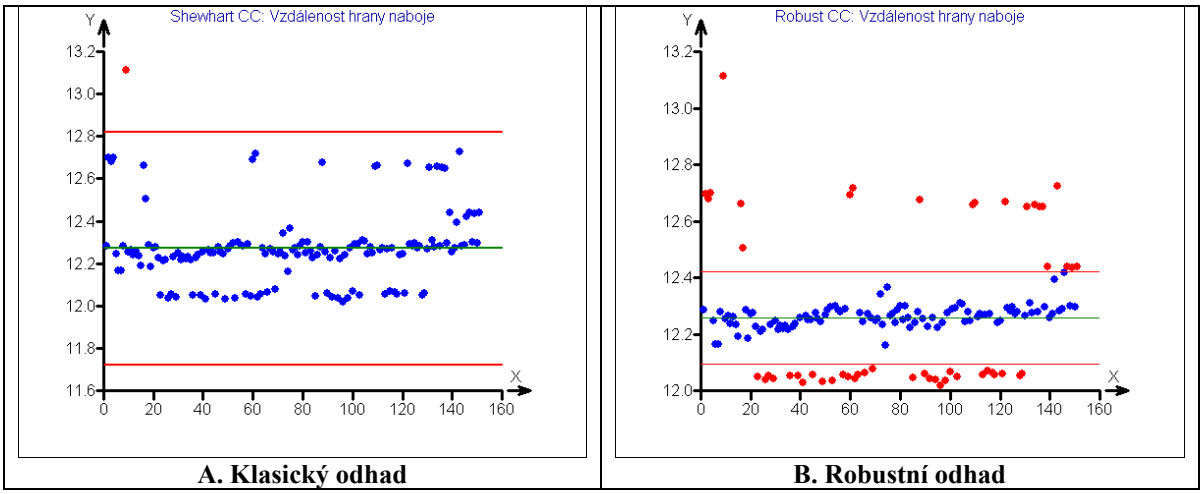
A. Klasický odhad **B. Robustní odhad**

Obr. 6 Odhad střední hodnoty a regulačních mezí klasickým a robustním postupem pro krytku



A. Klasický odhad **B. Robustní odhad**

Obr. 7 Odhad střední hodnoty a regulačních mezí klasickým a robustním postupem pro odlehčení



A. Klasický odhad **B. Robustní odhad**

Obr. 8 Odhad střední hodnoty a regulačních mezí klasickým a robustním postupem pro vzdálenost hrany

3.2. Další robustní metody

Užitečnost a častá oprávněnost použití robustních modelů vedla k jejich explozivnímu rozšíření a publikování od šedesátých let 20. století dodnes. Z nepřehledného množství metod a robustních odhadů polohy zmiňme alespoň ještě M-odhady a uřezaný průměr.

M-odhady [184][186], [188], [157], [181] minimalizují

$$\sum_{i=1}^{n} \rho\left(x_i - \mu\right) = \sum_{i=1}^{n} \rho\left(e_i\right). \qquad (9)$$

Použijeme-li za funkci $\rho = -\ln f(x)$, kde $f(x)$ je hustota pravděpodobnosti například normálního, nebo Laplaceova rozdělení, dostaneme maximálně věrohodné odhady μ ve

tvaru průměru a mediánu, jak bylo popsáno dříve. Položíme-li ψ(x) = ρ', přechází úloha na řešení rovnice $\sum_{i=1}^{n} \psi(e_i) = 0$. Definujeme-li dále

$$w(x) = \begin{cases} \psi(x)/x & pro \;\; x \neq 0 \\ \psi'(x) & pro \;\; x = 0 \end{cases}, \qquad \textbf{(10)}$$

pak lze chápat M-odhad polohy jako vážený průměr s váhami $w(x_i - \mu)$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w(e_i) x_i}{\sum_{i=1}^{n} w(e_i)}. \qquad \textbf{(11)}$$

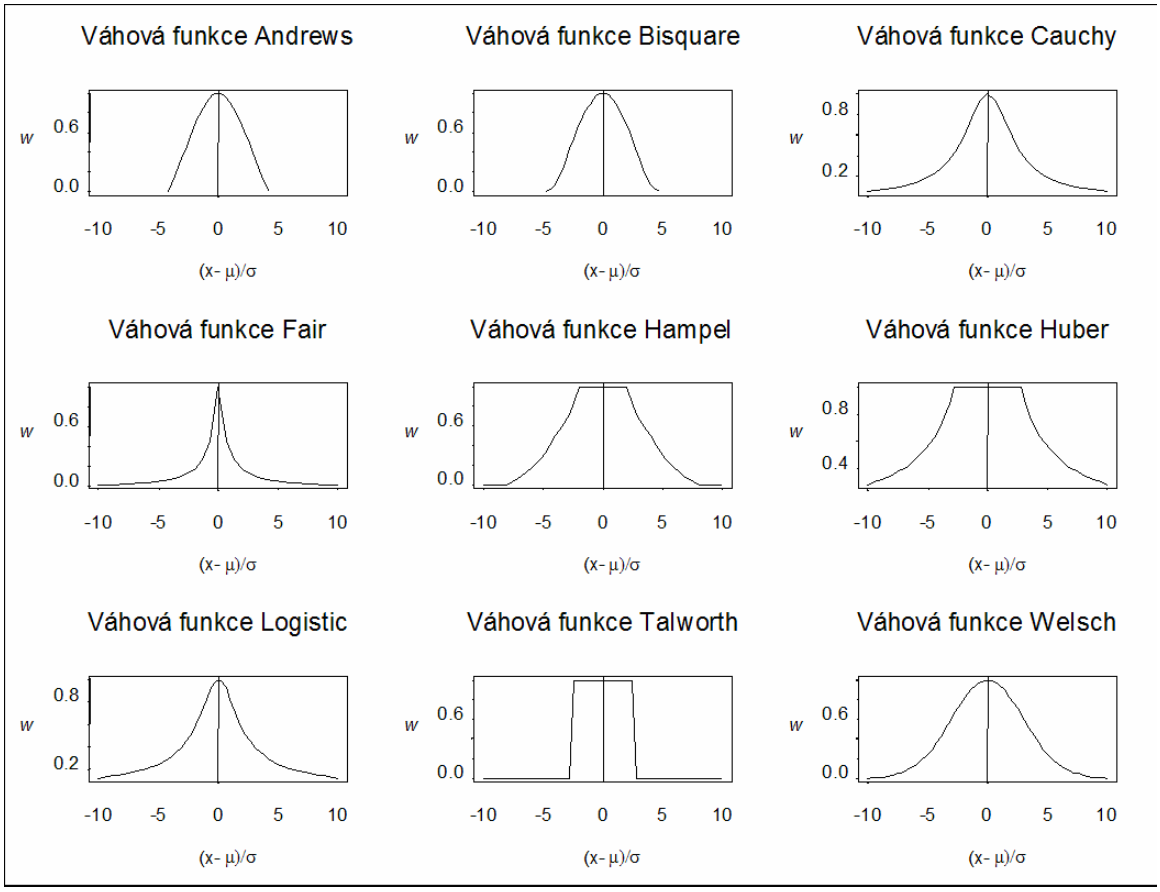
Například pro normální rozdělení a známé konstantní $\sigma^2 = 1$ je

$$f(x,\mu) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu)^2 \right],$$
$$\rho(x,\mu) = \frac{1}{2}\ln(2\pi) + \frac{1}{2}(x-\mu)^2,$$
$$\psi(x,\mu) = x - \mu$$

a váhová funkce $w(e) = 1$, kde $e = x - \mu$. Podobně pro Laplaceovo rozdělení je

$$\rho = |e|, \; \psi = \text{sign}(x)$$

s váhovou funkcí $w(e) = |1/x|$. Na rozdíl od normálního rozdělení je zde $w(e)$ klesající funkcí $|e|$, to snižuje váhu odlehlejších pozorování, důsledkem čehož je robustnost odhadu. Tvarem váhové funkce pro Laplaceovo rozdělení (a tedy robustní medián) je inspirována rodina robustních M-odhadů, kde jsou použity symetrické nezáporné váhové funkce $w(e)$ podobných tvarů s maximem v nule a neklesající na intervalu $(-\infty, 0\rangle$ a nerostoucí na intervalu $\langle 0, -\infty)$. Několik příkladů váhových funkcí normované odchylky e uvádí Obr. 9.



Obr. 9 Příklady grafů váhových funkcí pro M-odhady ve statistickém systému S-Plus

Pro všechny váhové funkce je charakteristické, že obsahují ladicí konstantu (tuning constant), jehož hodnota bývá doporučena autorem, nebo se nastavuje dle potřeby empiricky. Dále se zabýváme použitím M-odhadů v souvislostí s regresí v kapitole 5.

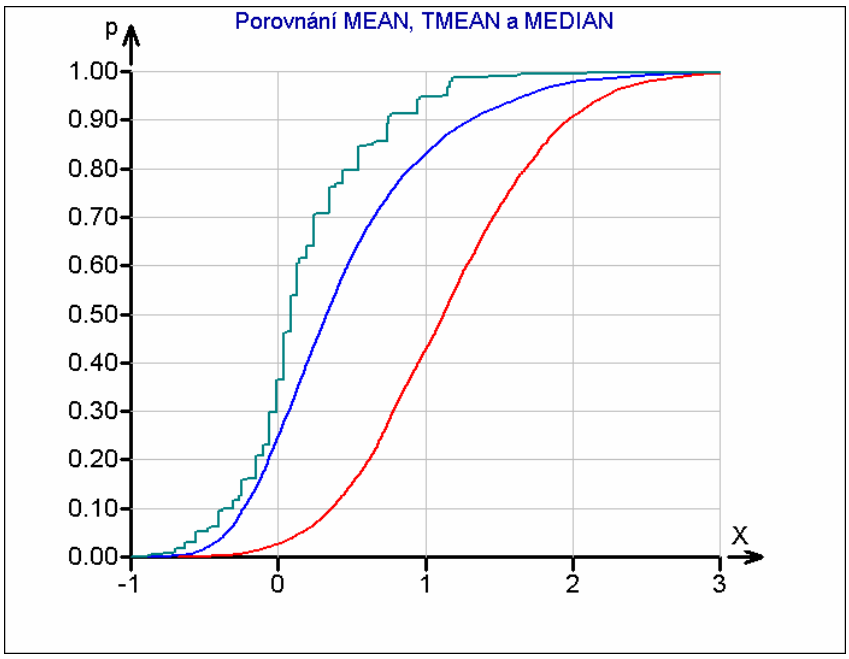
Uřezaný průměr, (např. [190], [198], [199], [196], [197], [202], [204], [205]) je počítán z výběru, z nějž byl odstraněn daný podíl α největších a nejmenších hodnot.

$$\overline{x}_\alpha = \frac{1}{n-2m}\sum_{i=m+1}^{n-m} x_{(i)}\,, \qquad \textbf{(12)}$$

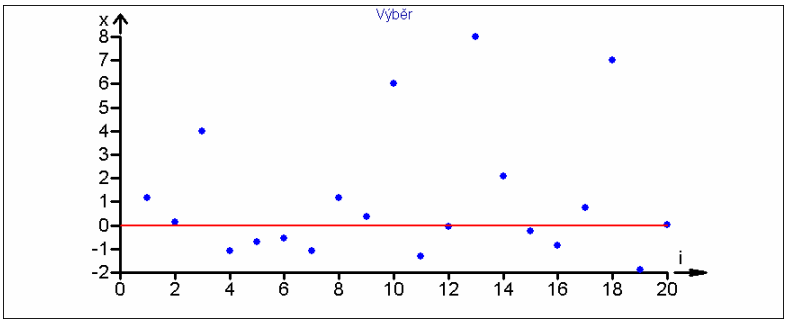
kde $x(i)$ je i-tý prvek setříděného výběru, $m = \mathrm{int}\left(\alpha\left(n-1\right)\right)$ a $\alpha \in \left\langle 0, 1/2\right)$ je podíl odstraněných hodnot. Tento často používaný robustní odhad patří do třídy L-odhadů založených na lineárních kombinacích uspořádaného výběru. Pro α = 0 dostáváme klasický průměr, pro $\alpha \to 0.5$ výběrový medián.

Na Obr. 10 jsou bootstrapem získané distribuční funkce klasického a 20% uřezaného průměru (α=0.2) a mediánu pro výběr z rozdělení $N(0,\ 1)$, x = (1.15, 0.13, 4, -1.084, -0.699, -0.559, -1.101, 1.186, 0.351, 6, -1.305, -0.055, 8, 2.091, -0.246, -0.872, 0.745, 7, -1.875, 0.038), viz Obr. 11, obsahující čtyři vybočující hodnoty 4, 6, 8 a 7. Odhadu μ = 0 se zřejmě blíží více medián a uřezaný průměr, než nerobustní klasický

průměr. Skript pro tento příklad je v Tab. 3. Navíc lze ukázat ([196], [204]), že odhad $\overline{x}_\alpha$ je efektivním odhadem střední hodnoty pro některá rozdělení s dlouhými konci, například Cauchyho rozdělení.



Obr. 10 Porovnání empirických distribučních funkcí aritmetického průměru (červeně), 20% uřezaného průměru (modře) a výběrového mediánu (zeleně) získaných pomocí bootstrapu výběru s rozsahem n=20 obsahujícího 4 vybočující hodnoty



Obr. 11 Výběr použitý v předchozím příkladě s vyznačenou střední hodnotou $\mu = 0$

Tab. 3 Skript pro bootstrap odhadů polohy

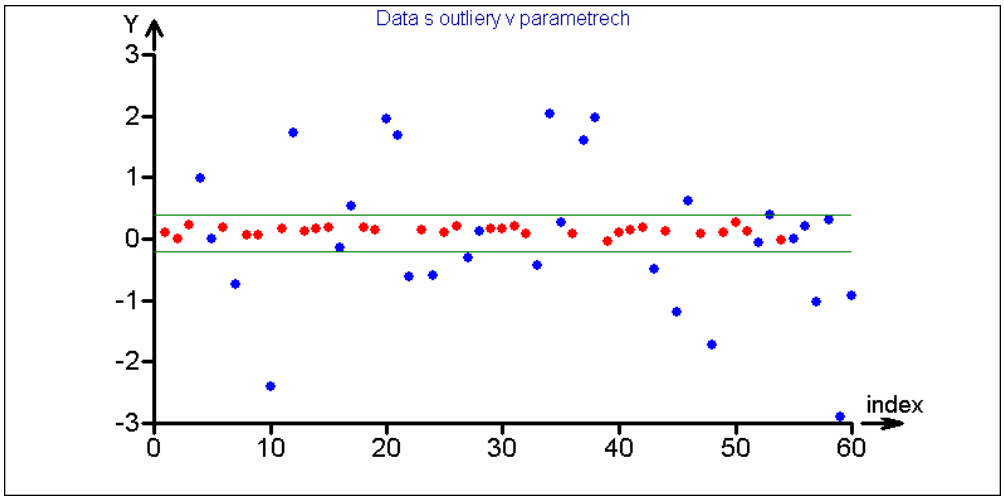
```
BN=5000
x=vec(normalr(16),8,6,4,7)
x=sample(x,20)
bs=bootstrap("tmean",x,BN)
bs1=bootstrap("average",x,BN)
bs2=bootstrap("median",x,BN)

plot(x,main="Výběr")
lineadd(h=0,color=3,width=2)
plot(sort(1,bs$bx),(1:BN)/(BN+1),type=line,width=2,main="Porovnání MEAN,
TMEAN a MEDIAN")
plotadd(sort(1,bs1$bx),(1:BN)/(BN+1),type=line,color=3,width=2)
plotadd(sort(1,bs2$bx),(1:BN)/(BN+1),type=line,color=5,width=2)
```

3.3. Směs normálních rozdělení

Pojem vybočující hodnoty v prostoru měřené veličiny X lze rozšířit na parametrický prostor statistického modelu. Zde by bylo vybočující takové měření, které pochází z rozdělení s odlišnými parametry (obecně z jiného rozdělení). Poskytuje-li tedy sledovaný fyzikální děj data s hustotou $f(\mathbf{x}, \boldsymbol{\theta})$, označíme jako vybočující každé měření pocházející z rozdělení s hustotou $f(\mathbf{x}, \boldsymbol{\theta}_1)$; $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}$. Vzálenost dvou naměřených hodnot $\mathbf{x}_i$, $\mathbf{x}_j$ lze pak definovat například jako

$$d_\theta = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}\|,$$

místo $d_x = \|\mathbf{x}_i - \mathbf{x}_j\|$. Je zřejmé, že takto definovaný jednotlivý outlier nelze obecně identifikovat ve výběrovém prostoru, jak je ilustrováno na Obr. 12, na němž je směs dat pocházejících z rozdělení $f_1 = N(0, 1)$ (modře) a $f_2 = N(0.1, 0.1)$ (červeně), použitý skript v jazyce DAR je v Tab. 4. Většina hodnot z intervalu řekněme (–0.2, 0.4) bude pocházet z f_2, avšak naopak 23.5% hodnot z f_1 budou rovněž ležet v témže intervalu. Tyto hodnoty nelze bez dalších informací od sebe jednotlivě rozlišit. Při znalosti parametrů f_1 a f_2 je však možné alespoň odhadnout počet hodnot pocházejících z jednotlivých rozdělení a tím tyto θ-outliery identifikovat.



Obr. 12 Data obsahující hodnoty vybočující v parametrech (červeně)

Tab. 4 Skript k Obr. 12

```
x=vec(normalr(30), normalr(30, mean=0.1, sdev=0.1))
i=order(1, random(1:60))
tp=vec(rep(0, 30), rep(3, 30))
plot(x[i], ptcolor=tp[i])
lineadd(h=vec(-0.2, 0.4), color=1) // -/+3sigma
```

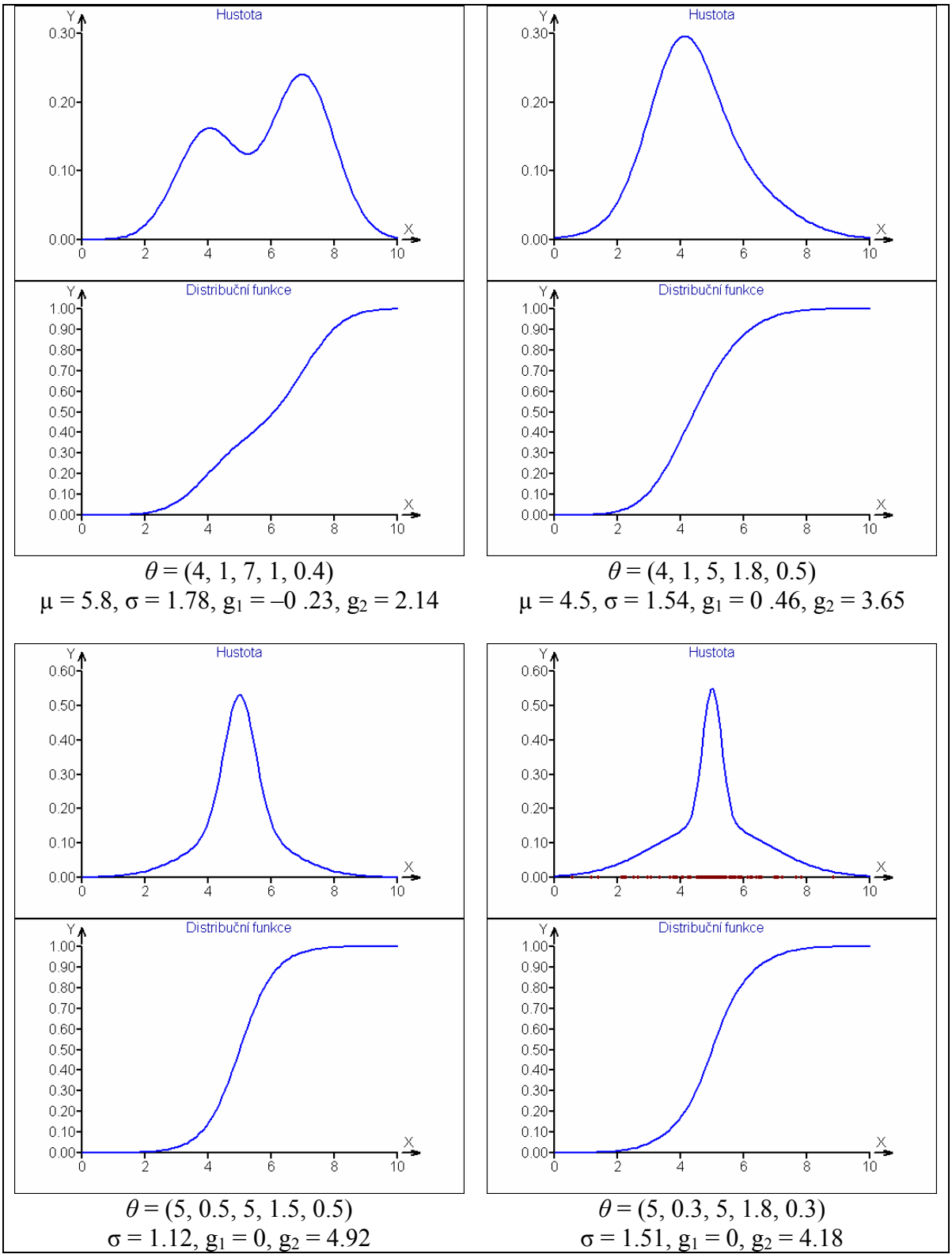
V praxi se velmi často vyskytuje situace, kdy jednorozměrný výběr pochází z několika, nejčastěji ze dvou fyzikálně odlišných zdrojů s odlišnými statistickými modely, kterými jsou výběry generovány. Například výrobky ze dvou linek, nebo od dvou různých

dodavatelů, parametry organismů ze dvou lokalit a podobně, přitom však není k dispozici informace o tom, ze kterého zdroje jednotlivá měření pocházejí, ani o tom, kolik hodnot pochází ze kterého zdroje. Úkolem je identifikovat parametry jednotlivých zdrojů, případně určit rozsahy dat z jednotlivých zdrojů. (Jedná se tedy obecně o úlohu typu „unsupervised learning“). Tato tématika je poměrně široce publikovaná s řadou aplikací, např. [62] - [74].V tomto odstavci se zaměříme na jednoduchý jednorozměrný případ směsi dvou normálních rozdělení (normal mixture), která může někdy přirozeně vysvětlit a interpretovat mimo jiné i data s velkou špičatostí, o nichž se zmiňoval předchozí odstavec.

Hustota rozdělení zde obsahuje pět parametrů $\boldsymbol{\theta} = (\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2)$ a lze ji zapsat ve tvaru

$$f(x;\alpha,\mu_1,\sigma_1,\mu_2,\sigma_2) = \frac{\alpha}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-\alpha}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right] \quad \textbf{(13)}$$

Podle hodnot jednotlivých parametrů lze získat rozdělení s různou šikmostí a špičatostí, jak ilustrují grafy hustoty a distribuční funkce pro některé hodnoty parametrů na následujících obrázcích.



$\theta = (4, 1, 7, 1, 0.4)$
$\mu = 5.8, \sigma = 1.78, g_1 = -0.23, g_2 = 2.14$

$\theta = (4, 1, 5, 1.8, 0.5)$
$\mu = 4.5, \sigma = 1.54, g_1 = 0.46, g_2 = 3.65$

$\theta = (5, 0.5, 5, 1.5, 0.5)$
$\sigma = 1.12, g_1 = 0, g_2 = 4.92$

$\theta = (5, 0.3, 5, 1.8, 0.3)$
$\sigma = 1.51, g_1 = 0, g_2 = 4.18$

Obr. 13 Tvary hustoty a distribuční funkce směsi dvou normálních rozdělení pro některé hodnoty parametrů

Pro odhad parametrů modelu (13) lze použít metodu maximální věrohodnosti, která maximalizuje výraz (14) vzhledem k parametrům.

$$\ln L(\mathbf{x};\theta) = \sum_{i=1}^{n} \ln \left\{ \frac{\alpha}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x_i-\mu_1)^2}{2\sigma_1^2} \right] + \frac{1-\alpha}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{(x_i-\mu_2)^2}{2\sigma_2^2} \right] \right\} \qquad \textbf{(14)}$$

Položíme-li parciální derivace rovny nule

$$\frac{\partial \ln L(\mathbf{x};\theta)}{\partial \theta'} = 0 \text{; kde } \theta = (\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2),$$

získáme soustavu pěti nelineárních rovnic [71]

$$\begin{aligned}
\frac{\partial \ln L}{\partial \alpha} &= \sum_{i=1}^{n} \left\{ \frac{1}{D} \left[\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2} \right) - \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2} \right) \right] \right\} &= 0 \\
\frac{\partial \ln L}{\partial \mu_1} &= \sum_{i=1}^{n} \left\{ \frac{1}{D} \left[\frac{(x_i-\mu_1)}{\sigma_1^2} \frac{\alpha}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2} \right) \right] \right\} &= 0 \\
\frac{\partial \ln L}{\partial \sigma_1} &= \sum_{i=1}^{n} \left\{ \frac{1}{D} \left[-\frac{\alpha}{2\sigma_1^3\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2} \right) + \frac{(x-\mu_1)^2}{2\sigma_1^4} \frac{\alpha}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2} \right) \right] \right\} &= 0 \qquad \textbf{(15)} \\
\frac{\partial \ln L}{\partial \mu_2} &= \sum_{i=1}^{n} \left\{ \frac{1}{D} \left[\frac{(x_i-\mu_2)}{\sigma_2^2} \frac{\alpha}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2} \right) \right] \right\} &= 0 \\
\frac{\partial \ln L}{\partial \sigma_2} &= \sum_{i=1}^{n} \left\{ \frac{1}{D} \left[-\frac{1-\alpha}{2\sigma_2^3\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2} \right) + \frac{(x-\mu_2)^2}{2\sigma_2^4} \frac{1-\alpha}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2} \right) \right] \right\} &= 0
\end{aligned}$$

kde $D = \frac{\alpha}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2} \right] + \frac{1-\alpha}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2} \right].$

Tuto úlohu lze řešit maximalizací (14), případně minimalizací $-\ln L(\mathbf{x};\theta)$ některou iterační optimalizační metodou, například Gauss-Newtonovým algoritmem.

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \mathbf{H}_i^{-1}\mathbf{g}_i \qquad \textbf{(16)}$$

V Lipschitzově okolí minima konverguje algoritmus kvadraticky, avšak ve větší vzdálenosti od minima klasický Gauss-Newtonův algoritmus konverguje pomalu, nebo diverguje z důvodu negativně definitního hessiánu. Tato nevýhoda může být zmírněna různými úpravami minimalizačního algoritmu. Zde byl navržen jednoduchý postup, při němž se rekonstruuje hessián funkce (14) pomocí vlastních čísel a vektorů, přičemž buď (1) případná záporná vlastní čísla jsou eliminována přičtením vhodné konstanty k $\boldsymbol{\Lambda}$, nebo (2) je v každém kroku posilována diagonála $\mathbf{H}$ přičtením $k\mathbf{I}$, k>0 tak, aby min($\boldsymbol{\Lambda}_{ii}$) > 0, jak naznačuje vztah (17).

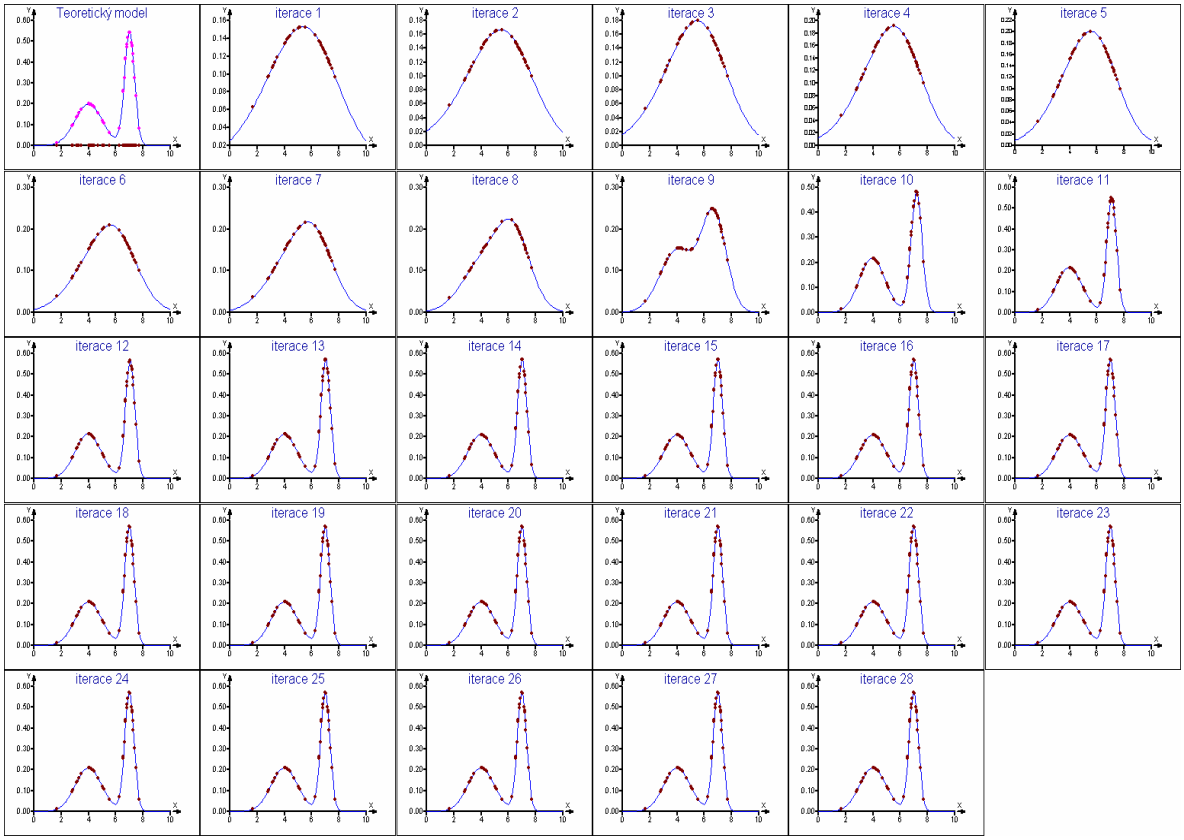
$$\mathbf{H} = \mathbf{Q}\boldsymbol{\Lambda}^{+}\mathbf{Q}^{T}, \qquad (17)$$

kde $\mathbf{Q}$ je matice vlastních vektorů hessiánu $\mathbf{H}$, $\boldsymbol{\Lambda}^{+}$ je diagonální matice s vlastními čísly na diagonále, kde záporná vlastní čísla byla eliminována přičtením kladné konstanty a $\mathbf{Q}^{T} = \mathbf{Q}^{-1}$ je transponovaná matice $\mathbf{Q}$. Tím se výrazně zlepšila konvergence Gauss-Newtonova algoritmu, který se tak stal použitelným pro odhadování parametrů $\boldsymbol{\theta}$ ve vztahu (14). Realizovaný algoritmus využívá první z výše zmíněných možností.

3.3.1. Implementace algoritmu

Následující ilustrace demonstruje práci realizovaného optimalizačního postupu, jehož kompletní výpis je uveden v Tab. 7. Dva příklady se simulovanými daty představují směs dvou normálních rozdělení s rozdílnými a blízkými středními hodnotami. Použité modelové hodnotamy parametrů byly v prvním případě $\mu_1 = 4$, $\sigma_1 = 1$, $\mu_2 = 7$, $\sigma_2 = \exp(-1)$, $\alpha = 1$, $n_1 = 20$, $n_2 = 20$, ve druhém případě $\mu_1 = 6.5$, $\sigma_1 = \exp(0.5)$, $\mu_2 = 7$, $\sigma_2 = \exp(-1)$, $\alpha = 1$, $n_1 = 30$, $n_2 = 70$. Průběh iterací je uveden v Tab. 5 a Tab. 6 a na Obr. 14 a Obr. 15.

Příklad 1

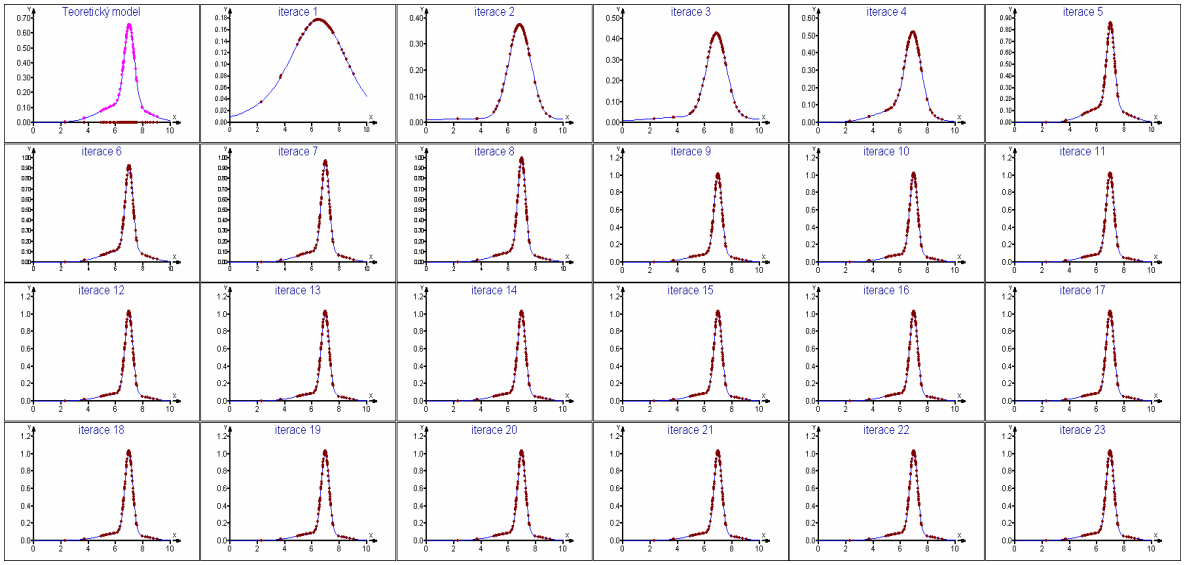


Obr. 14 Grafické znázornění postupných iterací pro případ $\mu_1 \neq \mu_2$

Tab. 5 Hodnoty parametrů, -log L, normy gradientu ||G||, tlumicího koeficientu optimalizace (dump), minimálního vlastního čísla hessiánu (min_eig) a počtu negativních vlastních čísel hessiánu (N_neg)

Iterace	m1	log(s1)	m2	log(s2)	atanh(alfa)	-logL	||G||	dump	min_eig	N_neg
0	4.0	1.0	6.0	1.0	0	86.4685	16.796	0.1		
1.0	3.9734	0.9882	5.9656	0.8021	-0.1752	83.4516	16.796	0.145	-5.1185	**1**
2.0	4.0258	0.9194	5.9549	0.7005	-0.2412	81.5291	14.9192	0.1653	-3.0209	**1**
3.0	4.1065	0.834	5.965	0.6253	-0.2649	80.0482	12.9594	0.1744	-2.1051	**1**
4.0	4.1852	0.754	5.9899	0.562	-0.2721	78.932	11.1123	0.1785	-1.401	**1**
5.0	4.2437	0.6857	6.028	0.5045	-0.272	78.1144	9.4404	0.1803	-0.838	**1**
6.0	4.2712	0.6246	6.0865	0.4454	-0.2632	77.4864	7.986	0.1811	-0.424	**1**
7.0	4.2434	0.5515	6.2041	0.3606	-0.2287	76.8185	6.7983	0.1815	-0.1855	**1**
8.0	4.1557	0.443	6.3909	0.2148	-0.1765	75.6175	6.1267	0.1817	-2.3539	**1**
9.0	4.0165	0.229	6.6964	-0.1027	-0.1037	71.5885	6.8881	0.1818	-7.076	**2**
10.0	3.9495	-0.063	7.1882	-0.908	0.0502	64.6321	11.2113	0.1818	-6.9926	**1**
11.0	3.9614	-0.0553	7.08	-1.0484	0.0475	62.4084	24.2771	0.2636	9.1704	0
12.0	3.9711	-0.049	7.0444	-1.0673	0.0426	61.9688	14.86	0.3373	9.2514	0
13.0	3.979	-0.0437	7.0198	-1.0716	0.0384	61.7739	9.6768	0.4035	9.3106	0
14.0	3.9849	-0.0397	7.0035	-1.0701	0.0353	61.6987	5.7537	0.4632	9.3573	0
15.0	3.9887	-0.037	6.9937	-1.0675	0.0335	61.6747	3.1056	0.5169	9.3867	0
16.0	3.9908	-0.0355	6.9885	-1.0654	0.0325	61.6686	1.5142	0.5652	9.4029	0
17.0	3.9919	-0.0347	6.986	-1.0643	0.032	61.6673	0.6646	0.6087	9.4108	0
18.0	3.9923	-0.0344	6.9849	-1.0638	0.0318	61.6671	0.2623	0.6478	9.4143	0
19.0	3.9925	-0.0343	6.9845	-1.0636	0.0317	61.6671	0.0931	0.683	9.4157	0
20.0	3.9926	-0.0342	6.9844	-1.0635	0.0316	61.6671	0.0297	0.7147	9.4162	0
21.0	3.9926	-0.0342	6.9844	-1.0635	0.0316	61.6671	0.0085	0.7432	9.4164	0
22.0	3.9926	-0.0342	6.9844	-1.0635	0.0316	61.6671	0.0022	0.7689	9.4164	0
23.0	3.9926	-0.0342	6.9844	-1.0635	0.0316	61.6671	0.0005	0.792	9.4164	0
24.0	3.9926	-0.0342	6.9844	-1.0635	0.0316	61.6671	0.0001	0.8128	9.4165	0
25.0	3.9926	-0.0342	6.9844	-1.0635	0.0316	61.6671	2.02772E-005	0.8315	9.4165	0
26.0	3.9926	-0.0342	6.9844	-1.0635	0.0316	61.6671	3.44356E-006	0.8484	9.4165	0
27.0	3.9926	-0.0342	6.9844	-1.0635	0.0316	61.6671	5.27103E-007	0.8635	9.4164	0
28.0	3.9926	-0.0342	6.9844	-1.0635	0.0316	61.6671	7.40798E-008	0.8772	9.4164	0
Final estimates:										
m1	**s1**	**m2**	**s2**	**N1**	**N2**	**-logL**				
3.9926	0.9664	6.9844	0.3453	20.3	19.7	61.6671				

Příklad 2



Obr. 15 Grafické znázornění postupných iterací pro případ podobných středních hodnot, μ_1= 6.5, μ_2=7

Tab. 6 Hodnoty parametrů, -log L, normy gradientu ||G||, tlumicího koeficientu optimalizace (dump), minimálního vlastního čísla hessiánu (min_eig) a počtu negativních vlastních čísel hessiánu (N_neg)

Iterace	m1	log(s1)	m2	log(s2)	atanh(alfa)	-logL	||G||	dump	min_eig	N_neg
0	6.0	1.0	6.5	1.0	0	200.0274	59.6675	0.1		
1.0	5.9437	1.1612	6.5533	0.598	-0.2589	183.6467	59.6675	0.145	-45.5471	**1**
2.0	5.757	1.9141	6.8196	-0.2075	-1.0083	139.6023	60.643	0.1653	-32.9999	**1**
3.0	5.7726	1.3024	6.8482	-0.3269	-0.9506	129.8227	40.1849	0.1744	-5.7859	**1**
4.0	5.9385	0.5627	6.9075	-0.5473	-0.6587	117.9373	36.0632	0.1785	-4.5221	**1**
5.0	6.623	0.3489	6.996	-1.2559	-0.0819	105.6042	31.5523	0.2606	0.6113	0
6.0	6.5664	0.3765	6.9816	-1.2637	-0.3081	104.0438	17.3794	0.3346	7.6731	0
7.0	6.5101	0.3916	6.973	-1.2659	-0.4681	103.3273	10.9813	0.4011	8.7918	0
8.0	6.4633	0.4007	6.968	-1.2654	-0.5782	103.0329	6.4081	0.461	9.1573	0
9.0	6.4302	0.4064	6.9652	-1.264	-0.6478	102.9314	3.4214	0.5149	9.1776	0
10.0	6.41	0.4098	6.9637	-1.2627	-0.6874	102.9029	1.6637	0.5634	9.0937	0
11.0	6.3994	0.4116	6.963	-1.2618	-0.7074	102.8964	0.7341	0.6071	9.0138	0
12.0	6.3945	0.4125	6.9627	-1.2614	-0.7165	102.8952	0.2933	0.6464	8.9645	0
13.0	6.3925	0.4129	6.9626	-1.2612	-0.7202	102.895	0.1059	0.6817	8.9403	0
14.0	6.3918	0.413	6.9626	-1.2611	-0.7215	102.895	0.0346	0.7135	8.9302	0
15.0	6.3915	0.4131	6.9625	-1.261	-0.7219	102.895	0.0102	0.7422	8.9264	0
16.0	6.3915	0.4131	6.9625	-1.261	-0.722	102.895	0.0027	0.768	8.9252	0
17.0	6.3914	0.4131	6.9625	-1.261	-0.7221	102.895	0.0007	0.7912	8.9249	0
18.0	6.3914	0.4131	6.9625	-1.261	-0.7221	102.895	0.0001	0.8121	8.9247	0
19.0	6.3914	0.4131	6.9625	-1.261	-0.7221	102.895	2.8E-05	0.8309	8.9248	0
20.0	6.3914	0.4131	6.9625	-1.261	-0.7221	102.895	5.0E-006	0.8478	8.9248	0
21.0	6.3914	0.4131	6.9625	-1.261	-0.7221	102.895	8.0E-007	0.863	8.9248	0
22.0	6.3914	0.4131	6.9625	-1.261	-0.7221	102.895	1.1E-007	0.8767	8.9247	0
23.0	6.3914	0.4131	6.9625	-1.261	-0.7221	102.895	1.6E-008	0.889	8.9247	0
Final estimates:										
m1	**s1**	**m2**	**s2**	**N1**	**N2**	**-logL**				
6.3914	1.5115	6.9625	0.2834	32.7	67.3	102.895				

3.3.2. Momenty směsi rozdělení

Momenty rozdělení (13) lze vyjádřit analyticky v podobě následujících vztahů (18) až (21).

$$\mu = \alpha\mu_1 + (1-\alpha)\mu_2 \qquad \textbf{(18)}$$

$$\sigma^2 = \alpha(1-\alpha)(\mu_1-\mu_2)^2 + \alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 \qquad \textbf{(19)}$$

$$g_1 = \frac{1}{\sigma^3}\alpha(1-\alpha)(\mu_2-\mu_1)\left[(2\alpha-1)(\mu_2-\mu_1)^2 + 3\left(\sigma_2^2-\sigma_1^2\right)\right] \qquad \textbf{(20)}$$

$$\begin{aligned}\sigma^4 g_2 = {} & -2\alpha^5(\mu_2-\mu_1)^4 + 6\alpha^4(\mu_2-\mu_1)^2\left[(\mu_2-\mu_1)^2 + \sigma_1^2 - \sigma_2^2\right] - \\ & -2\alpha^3(\mu_2-\mu_1)^2\left[2(\mu_2-\mu_1)^2 + 6\sigma_1^2 - 3\sigma_2^2\right] + \\ & +\alpha^2\left[3\left(\sigma_1^4-\sigma_2^4\right) + (\mu_2-\mu_1)^4 + 6\sigma_1^2(\mu_2-\mu_1)^2\right] + 3\alpha\sigma_2^4\end{aligned} \qquad \textbf{(21)}$$

3.3.3. Odhady pomocí moment vytvořující funkce

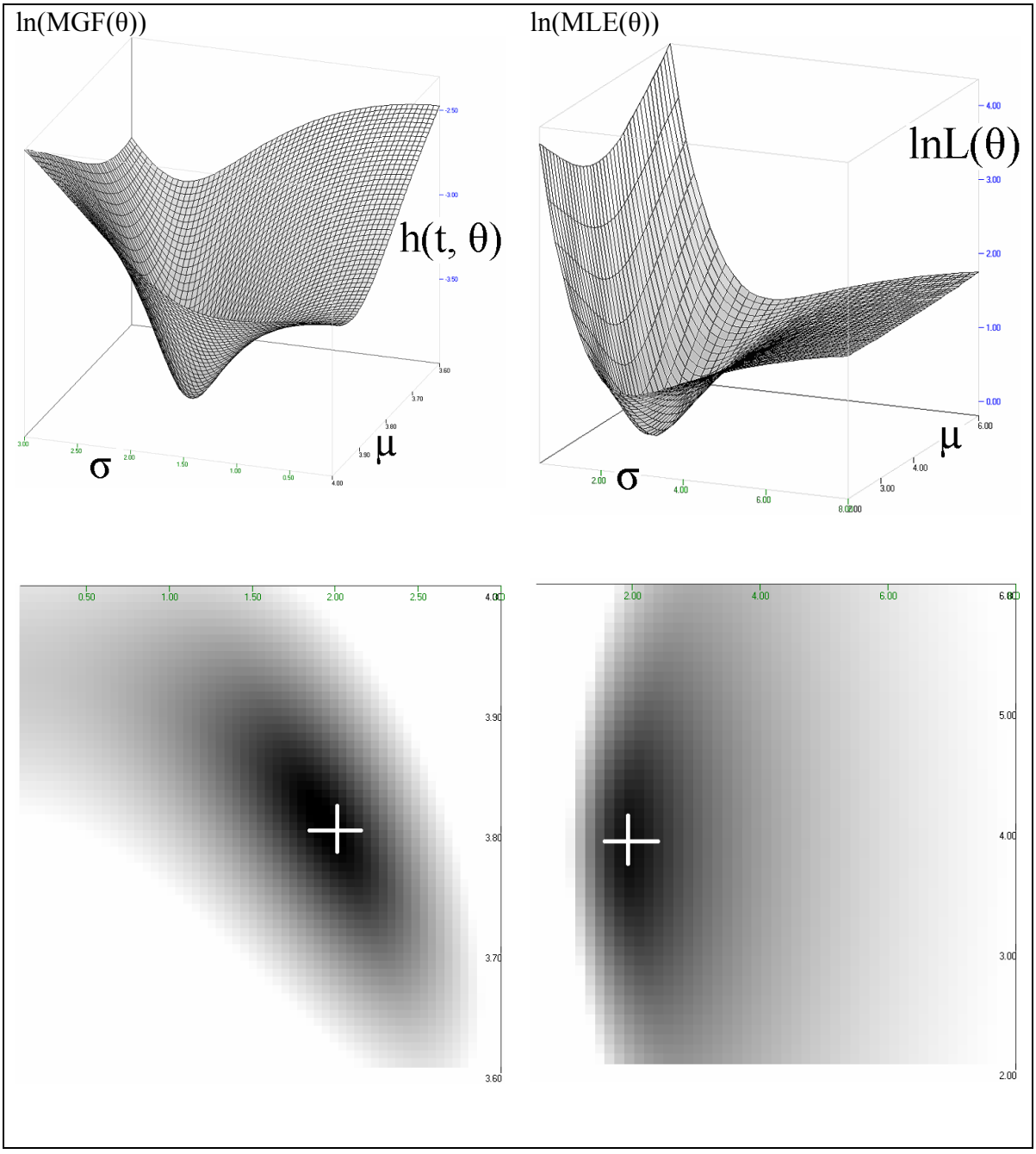
Alternativním postupem optimalizace parametrů hustoty (13) je minimalizace čtverců rozdílů mezi výběrovou a modelovou moment vytvořující funkcí (MGF, moment generating function, [71], [62]). Ta je definována jako střední hodnota

$$E\left(\mathrm{e}^{tX}\right)=\int_{-\infty}^{\infty}\exp(tx)\,f(x)\,\mathrm{d}x=(1-\alpha)\,\mathrm{e}^{\left(\frac{1}{2}t^2\sigma_1^2+t\mu_1\right)}+\alpha\,\mathrm{e}^{\left(\frac{1}{2}t^2\sigma_2^2+t\mu_2\right)}. \qquad \textbf{(22)}$$

Definujeme-li statistiku $g_s(t,x)=\frac{1}{n}\sum_{i=1}^{n}\exp(tx_i)$ jako funkci parametru t a výběru x, pak asymptoticky $g_s(t_j,x)\xrightarrow{p}g(t_j,\theta)$. Lze pak definovat vzdálenost (23), jejíž minimalizací pro m vhodně zvolených hodnot t_j lze získat optimální odhad parametrů θ. Lze ukázat, že stačí zvolit m rovné počtu parametrů, tedy v tomto případě m = 5. Volba hodnot parametru t je v podstatě libovolná, lze použít například hodnoty t = –2, –1, 0, 1, 2.

$$h(t,\theta)=\sum_{j=1}^{m}\left[g_s(t_j,x)-g(t_j,\theta)\right]^2. \qquad \textbf{(23)}$$

Quandt a Ramsey [73] uvádějí i asymptotické rozdělení takto získaného odhadu. Ve větší vzdálenosti od minima konverguje tato funkce k minimu lépe, než (14), avšak odhady parametrů podle (23) jsou závislé, jak ilustrují i tvary obou funkcí na Obr. 16. Tato ilustrace představuje tvar minimalizované funkce (14) a (23) pro dva parametry θ = (μ, σ) s polohou minima θ_{opt} = (3.8, 2.1) pro výběr **x** = (1,3,4,4,7). Určitou nevýhodou tohoto přístupu je závislost mezi μ a σ v okolí minima MGF, která nenastává v případě MLE. Minimalizace záporného logaritmu věrohodnosti –log L(θ) je proto obvykle snažší a byla použita i v algoritmu v Tab. 7 v odst. 3.3.4. Tvary kritérií MGF i MLE jsou ilustrovány a porovnány na následujícím Obr. 16.



Obr. 16 Porovnání tvaru MGF-vzdálenosti a věrohodnosti v okolí minima, poloha křížku vyznačujícího minimum je pouze orientační

3.3.4. Optimalizační algoritmus

Algoritmus výpočtu maximálně věrohodných odhadů neznámých parametrů hustoty (13) pomocí derivační modifikované Gauss-Newtonovy optimalizace je uveden v Tab. 7. Je napsán a odladěn v programovém prostředí DARWin, jehož popis a definice je v příloze této práce.

Tab. 7 Algoritmus použitý pro odhad parametrů směsi rozdělení

```
a=vec(5,0,5,-1,0)
a0=a
nn1=15;nn2=30
n0=nn1+nn2
a1=a[1];a2=a[2];a3=a[3];a4=a[4];a5=a[5]
b5=1/(1+exp(-a5))
b2=exp(a2);b4=exp(a4)

rx=normalr(nn1,mean=a1,sdev=b2)
rx=vec(rx,normalr(nn2,mean=a3,sdev=b4))
n=nrows(rx)

x1=seq(0,10,count=100)
p1=seq(0.001,0.999,count=100)
plot(x1,dens2(x1,a),type=LINE,main="")
plotadd(rx,dens2(rx,a),color=8)
plotadd(rx,seq(0,0,count=n),color=8)
plot(x1,distr2(x1,a),type=LINE,main="")
plotadd(rx,distr2(rx,a),color=8)
a=vec(4,1,6,1,0) // Počáteční odhady parametrů
dump=0.1 // Tlumicí koeficient
lik=loglike(rx,a) // Věrohodnost
gra=gradient(a,rx)
granorm=sqrt(sum(gra*gra)) // Norma gradientu
print("Index",\t,"m1",\t,"log(s1)",\t,"m2",\t,"log(s2)",\t,"atanh(alfa)",
\t,"-logL",\t,"||G||",\t,"dump",\t,"min_eig",\t,"N_neg",\n)
print(0,transp(a),lik,\t,granorm,\t,dump,\n)

graphsheet(cols=6)
plot(x1,dens2(x1,a0),type=LINE,main="Teoretický model")
plotadd(rx,dens2(rx,a0),color=7)
plotadd(rx,seq(0,0,count=n),color=8)

ii=0
while(GT(granorm,1e-7))
{
gra=gradient(a,rx)
hes=hessian(a,rx)
//eigval=sort(1,eigen1(hes))
eigval=eigenval(hes)
eigvals=sort(1,eigval)
eigvec=eigenvec(hes)
eigmin=eigvals[1] // Minimální vlastní číslo
eigmin0=eigmin
nnegatives=-sum((sign(eigval)-1)/2) // Počet negativních vlastních čísel
hessiánu
// Zajištění pozitivní definitnosti hessiánu (1):
eva=(eigval-eigmin)+1 // Posunutí vlastních čísel, aby min(L)=1
if (LE(eigmin,0)) // Pokud je minimální vlastní číslo záporné, ...
{
hes=eigvec#diag(eva)#transp(eigvec) // rekonstrukce hessiánu QLQ
dump=dump/2 // a zmenšení tlumicího koeficientu
}
// Alternativní zajištění pozitivní definitnosti hessiánu posílením
diagonály (2):
/*
ei=0
while(lt(eigmin,0))
{
hes=hes+1*unit(5)
```

```
eigmin=sort(1,eigen1(hes))
eigmin=eigmin[1]
dump=dump/2
ei=ei+1
//eigmin
}
*/
dir=-pinv(hes+(1-dump)*unit(5))#gra // Kvazi-Newtonův krok d=H^(-1)*g
dump=1-(1-dump)*0.9 // zvětšení tlumicího koeficientu
a=a+dump*dir // provedení kroku
lik=loglike(rx,a)
granorm=sqrt(sum(gra^2))
ii=ii+1
print(ii,transp(a),lik,\t,granorm,\t,dump,\t,eigmin0,\t,nnegatives,\n)
//print(bind(a,gra),\t,lik,\t,dump,\t,ei,\t,granorm)
plot(x1,dens2(x1,a),type=LINE,main="iterace "+ii)
plotadd(rx,dens2(rx,a),color=8)
//************************
}

print(\n,"Final estimates:",\n)
print(\n,"m1",\t,"s1",\t,"m2",\t,"s2",\t,"N1",\t,"N2",\t,"-logL",\n)
n01=round(n0*1/(1+exp(-a[5])),1);n02=n0-n01
print(round(a[1],5),\t,round(exp(a[2]),5),\t,round(a[3],5),\t,round(exp(a
[4]),5),\t,n01,\t,n02,\t,round(lik,5),\n)

// ******************************

// Směs dvou normálních rozdělení a5*N(a1,a2)+(1-a5)*N(a3,a4)

function loglike1(x,a)
//Standardní tvar N(mu,sigma^2)
{
pi=3.14159265358979
a1=a[1];a2=a[2];a3=a[3];a4=a[4];a5=a[5]
ss1=1/sqrt(2*pi)/a2*exp(-((x-a1)^2)/(2*a2*a2))
ss2=1/sqrt(2*pi)/a4*exp(-((x-a3)^2)/(2*a4*a4))
ss=sum(ln(a5*ss1+(1-a5)*ss2))
return(-ss)
}

function loglike(x,a)
// Exponenciální tvar N(mu,exp(log(sigma))^2);
// tanh(alfa)
{
pi=3.14159265358979
a1=a[1];a2=a[2];a3=a[3];a4=a[4];a5=a[5]
b5=1/(1+exp(-a5))
b2=exp(a2);b4=exp(a4)
ss1=1/sqrt(2*pi)/b2*exp(-((x-a1)^2)/(2*b2*b2))
ss2=1/sqrt(2*pi)/b4*exp(-((x-a3)^2)/(2*b4*b4))
ss=sum(ln(b5*ss1+(1-b5)*ss2))
return(-ss)
}

//**************************************

function dens02(x,a)
//Standardní tvar N(mu,sigma^2)
{
dd=x
```

```
n=nrows(x)

b5=1/(1+exp(a[5]))
dd1=1/sqrt(2*pi)/a[2]*exp(-((x-a[1])^2)/(2*a[2]*a[2]))
dd2=1/sqrt(2*pi)/a[4]*exp(-((x-a[3])^2)/(2*a[4]*a[4]))
dd=a[5]*dd1+(1-a[5])*dd2
return(dd)
}
function dens2(x,a)
//Exponenciální tvar N(mu,exp(log(sigma))^2)
// tanh(alfa)
{
dd=x
n=nrows(x)
a1=a[1];a2=a[2];a3=a[3];a4=a[4];a5=a[5]
b5=1/(1+exp(-a5))
b2=exp(a2);b4=exp(a4)
dd1=1/sqrt(2*pi)/b2*exp(-((x-a1)^2)/(2*b2*b2))
dd2=1/sqrt(2*pi)/b4*exp(-((x-a3)^2)/(2*b4*b4))
dd=b5*dd1+(1-b5)*dd2
return(dd)
}

function distr2(x,a)
//Exponenciální tvar N(mu,exp(log(sigma))^2)
// tanh(alfa)
{
dd=x
a1=a[1];a2=a[2];a3=a[3];a4=a[4];a5=a[5]
b5=1/(1+exp(-a5))
b2=exp(a2);b4=exp(a4)
n=nrows(x)
dd1=normalp((x-a[1])/sqrt(b2))
dd2=normalp((x-a[3])/sqrt(b4))
dd=b5*dd1+(1-b5)*dd2
return(dd)
}



**************************************************

function gradient(x,t)
// a --- parametry, t --- data
{
b=x
k=nrows(x)
eps=0.00001
z=seq(0,0,count=k)
for (i=1,k)
{
b[i]=b[i]-eps
z1=loglike(t,b)
b[i]=b[i]+2*eps
z2=loglike(t,b)
b[i]=b[i]-eps
z[i]=(z2-z1)/(2*eps)
}
return(z)
}
```

```
//*****************************************************

function hessian(x,t)
// a --- parametry, t --- data
{
b=x
k=nrows(x)
eps1=0.00001
eps2=0.0001
z=unit(k)
for (i=1,k)
{
for(j=1,k)
{
b[j]=b[j]-eps2
b[i]=b[i]-eps1
z1=loglike(t,b)
b[i]=b[i]+2*eps2
z2=loglike(t,b)
b[i]=b[i]-eps2
zz1=(z2-z1)/(2*eps2)

b[j]=b[j]+2*eps1
b[i]=b[i]-eps2
z1=loglike(t,b)
b[i]=b[i]+2*eps2
z2=loglike(t,b)
b[i]=b[i]-eps2
zz2=(z2-z1)/(2*eps2)
b[j]=b[j]-eps1
z[i,j]=(zz2-zz1)/(2*eps1)
}
}
return(z)
}
```

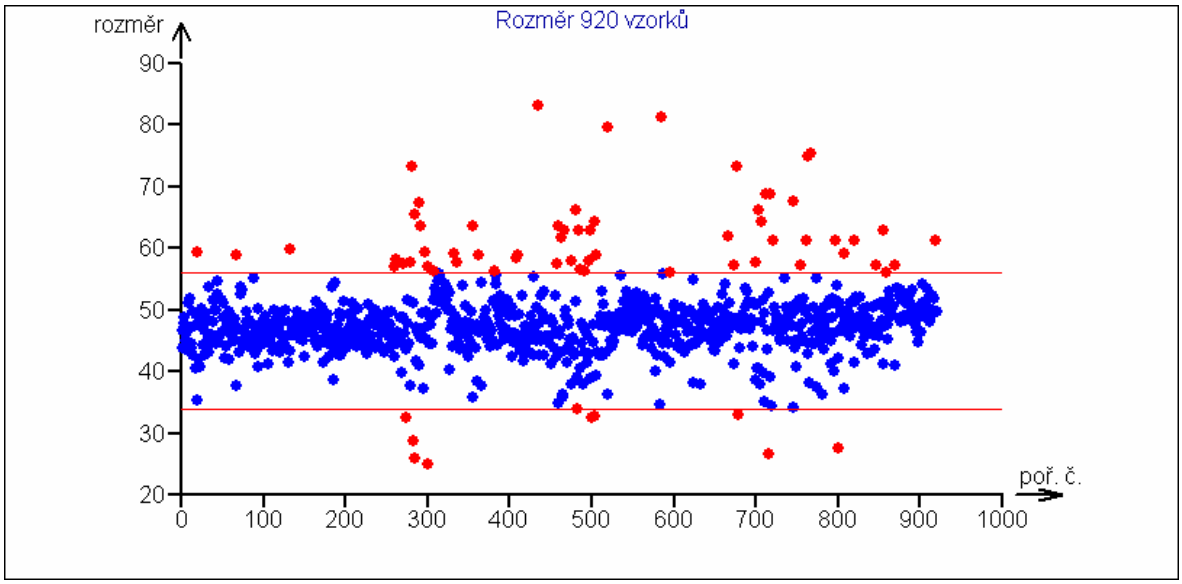
3.3.5. Aplikace metody

Příklad 1

Data (Obr. 17) představují změřený kritický rozměr zahraničního polotvaru pro ražbu českých mincí při vstupní přejímce do mincovny ČNB v relativních jednotkách. Přípustné hodnoty jsou mezi 34 a 56. Mimo tyto meze je mince považována za neplatnou a odmítána platebními automaty. Rozdělení tohoto rozměru je obecně důvodně považováno za normální. Výběr 920 vzorků obsahuje 70 nevyhovujících hodnot ale neodpovídá dobře žádnému z běžných rozdělení (podle K-S testu shody rozdělení na hladině α=0.05). Data jsou úmyslně lineárně transformována a uvedena bez jednotek, aby mohl být tento příklad zveřejněn.

46.7	45.9	50.1	46.4	38.7	44.7	44.5	51.7	48.7	52.4	46.7	43.1	48.7	49.1	48.1	47.2	50.3	43.5	44.3	48.6
43.9	46	40.8	47	45.5	45.7	44.4	53.6	47.4	44.7	45.9	47.5	52.8	50.9	46.5	46.9	45.1	48.1	50.1	48.1
47.8	51.5	45.6	49.6	43.6	50.2	37.7	49.9	50	41.6	61.6	42.7	47.2	50.4	45	51	49.1	46.1	47	53.3
48.8	46.3	44.4	48.7	54.5	48.1	57.6	52.8	45.2	46.4	35.8	48.9	47.9	47.9	43.5	48.6	48	48.5	45.5	51.5
46.6	48.8	45	48	43.7	47.5	73.2	40.2	48	47.3	36.3	42.7	50.4	45.9	49.7	47.6	48.7	46.4	48.9	49.1
45	51.2	44.6	47.3	47.6	43.9	28.8	46.8	46.7	47.4	62.9	52.9	49.4	44.1	47.8	44.1	50.4	46.4	48.1	48.4
46.4	42.1	48	45.2	46.8	50	51.2	50	47.2	44.6	43.4	43.1	49.6	48.2	43.4	51	49.3	51.5	42.9	48.2
43.3	47.1	44.4	48.2	49.4	49.5	26	44	44	46.8	49.1	50.8	51.6	44.3	46.1	50.9	47.3	50.8	50.3	52.5
51	44	46	47.6	42.6	48.1	65.5	48	47.2	47.2	50	42.7	48.6	48.2	48	57.8	67.6	45	48.5	52.1
43.7	47.5	48.4	46.1	42.6	47.3	41.6	47.3	46.6	43.1	41.5	47.3	44.2	44.6	50.4	38.7	34.2	41.2	46.5	51
51.8	46.9	49	44.9	48.8	44.5	47	59.2	51.3	45.9	47.5	48.2	45.3	50.5	47	50.8	43.9	51.9	46.6	50.8
43.7	41.9	47.1	49.1	44.7	48.2	46.4	49	45.1	49	48.2	43.1	52.2	43.5	44.8	40.6	45	48.8	47.1	52.4
44.4	49.7	47.5	42.5	45.9	46.2	40.9	57.7	56.2	47	45.3	36.3	51.3	43.9	50.2	66.1	50.8	40	43	52.4
49.6	49	41.3	46.6	51.3	47.9	67.4	48	51.7	46.5	44.4	79.7	50.1	47.7	47.8	40.5	40.7	61.2	50.3	49.6
42.6	45.7	46.6	48.3	46.3	49.1	45.4	44.5	55.7	55.4	57.9	47.6	45.9	43	52.5	37.9	50.1	51.7	45.9	50.8
47.4	46.1	45.1	46.4	45.1	46.5	63.5	45.9	54.1	42.6	38	48.9	50.5	49.5	48.4	64.4	47.1	46.3	50.7	53.3
49.2	45	47.5	48.1	43.8	48.7	46.6	43.5	44.3	47.5	45.4	44.9	48	47.7	49.6	48.1	51.6	53.9	43.6	49.1
40.6	47.6	45.9	48.4	45.3	43.2	49.4	44.2	50.4	45	43.1	47.1	50.1	46.4	54.1	39.9	45	27.5	57.2	48.2
59.3	44.9	43.3	48.9	45.6	46.5	37.3	47.6	48.9	50.3	42.6	44.1	49.1	49.7	46.5	51.5	57.2	42.2	46.3	49.7
35.3	48.5	46.2	48.4	49.3	49.8	59.3	53.9	47.8	46	51.3	44.8	49.1	44	45.6	49.5	45.8	50.3	47.3	47.6
42	59	49.9	47.5	44.2	43.9	47.2	43.8	52.3	83.1	66.3	44.9	43.8	47	45.6	35	47.8	51.1	45.8	46.4
47.8	37.6	43.7	48.6	49.4	47.9	49.8	47.9	49.7	42.7	33.9	47.6	47	44.3	62	68.8	51.8	49.5	46.9	50
40.8	46.8	46.9	49.5	47.5	45.7	45.2	46.3	47.3	47.7	39	46.4	45.5	49.9	46.8	43.5	49.7	47	51	44.7
50.1	46.8	49.7	45.1	48.5	43.8	24.9	50.7	46.2	52.2	62.8	48.4	46.6	44.5	51	46.6	50.1	47.4	47.1	49.6
49.5	43.1	48.8	45.6	51.2	47.6	57.1	48	50.8	45.1	56.6	47.2	40.1	38.2	48.7	26.5	61.3	37.1	47.1	48.7
44.3	53.7	46.8	47.3	45.2	46.2	44.5	43.3	45.2	49.5	40.5	48.1	51.9	54.9	48.4	52.8	44.8	59.1	48.2	50.1
49.2	52.5	46.2	51.4	46.5	49.7	49.1	45.4	50	53	47.7	46.7	45.6	49	47.5	68.9	42.8	48.2	41.1	46.3
46.7	44.8	47.5	44	46.1	45.5	47.6	42.4	46.8	42.6	44.6	49.4	42.6	46.5	47.4	39.2	75	46.4	63	54.1
42.7	53.3	43	48.7	46.2	57.1	49.1	44.3	45.6	45.4	37.9	55.5	49.7	44.5	41.3	34.3	38.2	49.6	50.1	51.2
49.5	45	48.5	47.6	44	42.3	44.4	47.8	51.6	45.8	47.3	48.4	50.1	48.6	57.2	61.2	75.3	49.7	56.1	50.8
45.1	46.3	44.8	48.5	45.2	58.1	56.2	49.7	48.9	45.8	43	48.3	34.7	45.3	48.5	46.6	48.6	47.8	51.3	50.7
48.4	46.1	44.3	47.9	50.1	44.6	51.6	44.5	46.4	48.6	56.2	51.3	81.3	49.3	47.3	47	46.5	51.2	45.3	50.2
43	48.1	45.6	50.4	46.6	44.4	52	35.9	49.3	45.3	46.5	48.3	49.2	48.7	73.2	50.5	44.4	45.7	47.5	49.4
53.6	49.7	45.4	44.9	47.5	46.5	53.5	63.7	42.3	44.4	41.5	49.7	44.5	38	32.9	47.5	44.6	47.3	51.5	48.9
44.8	44	49.9	41.5	45.3	44.8	49.4	46.9	50.5	45.7	45.9	48.3	55.8	46.8	48.4	49.3	50.5	48.5	53.1	53.5
45.7	47.5	44.1	46.8	45.8	48.7	54.6	47.2	45.1	41.3	42.8	46.7	42.4	48.4	48	47	44.4	46.9	48.9	51.8
46	48.4	45.8	47.3	46.7	45	51.3	43.8	44.7	48.5	57.9	53.1	45.7	45.5	43.8	48.2	55.2	61.3	49	50.2
51.5	43.9	48.1	44.7	46.5	44.6	55.6	47.6	46.1	48	38.8	48.1	46.9	50.8	46.9	48.2	37.5	41.4	49.3	48.1
48.1	45	41.5	48.3	45.4	39.7	55.8	38.4	48.9	47.7	62.8	49.5	49	50.3	48.7	45.5	47.3	52.1	50.6	49.1
45.8	44.9	59.8	49.3	47.6	57.5	52.3	59	45.3	43.9	32.4	51.1	48	45.6	49.3	45	48.8	51.3	50.7	49
49.9	47.6	43.3	44.6	43.9	48.8	54.8	46.5	58.5	46.2	44.7	52	44.6	45.9	47.8	47.8	46.1	48.3	40.9	52.2
50	46	46.5	46	46.4	46.1	53	46.4	59	52.4	47.6	49.3	46.2	45.6	47.5	47.1	47.4	48.7	57.3	50
44.6	55	45.9	42.4	42.8	32.5	51.6	37.6	49.2	42.3	64.3	50.6	41.4	48.5	50.5	46.6	49.8	47.4	53.5	51.9
54.7	42.9	46.6	46.8	50.8	43.6	52.8	54.4	44	57.5	32.7	46.4	56.1	49.4	53.5	47.3	44.8	52.1	53.4	49.8
52.6	46.6	46.8	53.7	46	51.6	54.1	47.6	43.4	34.8	39.3	48.7	47.5	49.2	52.9	55.1	36.3	47	53.5	61.2
44.8	43.3	45.9	45.5	48.5	47.7	50.6	46.6	43.9	63.6	59	48.3	48.6	44.9	51.3	50.3	50.4	48.8	52.2	49.8

Obr. 17 Naměřená data z ČNB



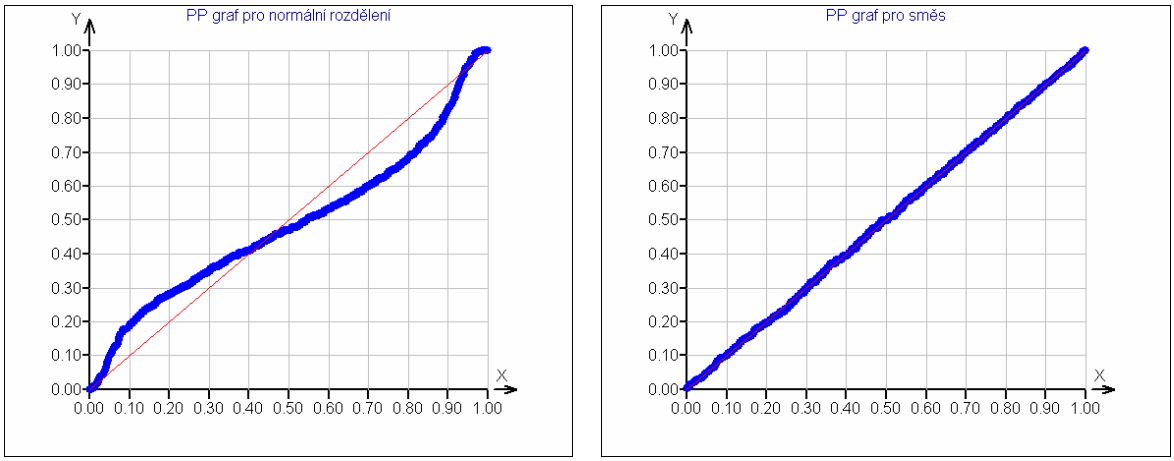
Obr. 18 Grafické znázornění dat s vyznačenou technologickou tolerancí

Pomocí popsané metody byly odhadnuty parametry směsi dvou normálních rozdělení s výsledkem uvedeným v Tab. 8 a s výbornou shodou s daty podle K-S testu dobré shody rozdělení. Shoda modelu s daty je patrná také porovnáním grafů (A) a (B) na Obr. 19

Tab. 8 Maximálně věrohodné odhady parametrů směsi 2 normálních rozdělení

μ_1	σ_1	μ_2	σ_2	N_1	N_2	-log$L$
47.32222	2.77439	49.80155	10.34495	699.9	220.1	2729.22035

Z těchto odhadů vyplývá, že 24% dodávky pochází z rozdělení se 14x větším rozptylem, což by naznačovalo, že čtvrtina dodávky je vyrobena chybnou, nestandardní a smluvně nepodloženou technologií. To se nakonec potvrdilo, dodavatel uznal svou chybu a český stát uspěl s rozsáhlou reklamací bez arbitráže.



(A) P-P graf pro normální rozdělení **(B) P-P graf pro směs dvou normálních rozdělení**

Obr. 19 Porovnání shody dat s modelem

Příklad 2

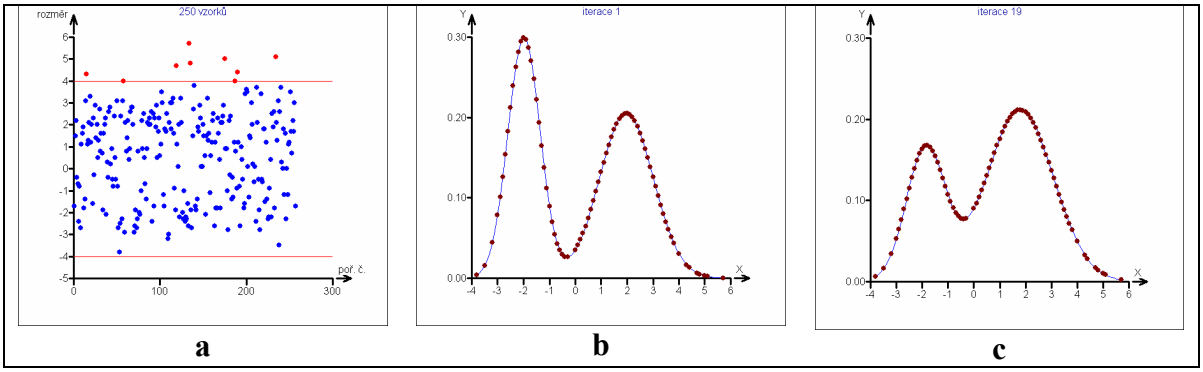
V Tab. 9 a na Obr. 20 jsou uvedena naměřená data z namátkově vybraných odlitků kol pro osobní automobily z podniku Hayes-Lemmerz. Data představují odchylky od nominální vzdálenosti dvou zvolených bodů odlitku v setinách milimetru. Odlitky pocházejí v poměru počtu 1:2 z dvou procesních cest, které není možno rozlišit a existovalo podezření posílené grafickým zobrazením dat, že obě cesty mohou dávat odlišné výsledky. Data jsou úmyslně uvedena bez jednotek a bližší specifikace.

Tab. 9 Tabulka naměřených dat z Hayes-Lemmerz

-1.7	2	-3.8	2.2	-1.7	-2.3	1.6	-1.4	-0.6	2.1
1.5	0.5	-2.5	0.9	1.8	-1.6	-1.1	-0.4	0.1	1.3
2.2	1.3	2.4	2.5	1.7	-2.6	0.6	-1.3	-2.4	3.1
-0.4	2.7	-2.3	0.5	2.5	5.7	1.2	1.2	3.7	-3.5
-0.7	0.8	3.1	-0.7	-3.2	4.8	2.2	4	2.1	2.6
-2.4	1.6	4	-1.4	-3	-1.9	2.5	0.8	-0.5	3.7
-0.8	0.7	-2.9	0.9	2.1	-2.7	-2.6	-0.9	-1.9	0.3
-2.7	-1.9	2.1	2.1	3	1.5	1.2	4.4	-2.8	0

1.1	1.6	0.8	2	1.8	2.7	-1.8	1.2	3.4	1.8
1.6	2	1.4	2.3	3	3.8	-1.8	-0.8	-0.5	-1.2
1.9	2.3	2.2	2	3.2	-1.7	2.3	-2.6	-0.6	1.6
-1.8	-2.1	0.4	2.6	1.6	-1.9	1.6	1	1.8	-2.3
-1.4	0.3	0.9	-2.4	-0.9	-1.4	0.4	0.4	1.8	0
3.1	-0.4	2.6	2.5	-1.9	1.9	-1.8	-0.5	-1.5	-1.2
4.3	2.6	2.8	-1.7	4.7	1.7	2.1	-1.8	1.2	1.1
1.1	2.8	2.8	2.3	1.1	2.9	2.1	-1.6	-1.4	1.2
1.3	0.2	2	1.9	0.1	2.3	2.4	3.4	-2.3	3.5
2.1	0.9	-2.9	2.9	2	0.1	2.9	3.6	-2.2	1.7
3.3	-0.5	-2.6	0.5	-2.2	0.1	5	3.5	-0.9	0.7
1.2	-0.8	-1.9	1.7	3.2	2	0.9	0.1	0.4	2.2
2	2.4	-2.7	2.3	2.1	1.5	0.9	-1.3	-2.2	1.7
-1.6	0.5	-2.3	0.6	-2	0.6	-1	1.5	2.5	3
2.3	-0.5	0.8	1.1	-2.3	-2.1	-2.8	1.4	1.9	-1.7
-0.3	3.1	-2	3.5	-1	1.5	2.2	3	2.6	
2.9	-0.8	-2.1	3	-2.4	3.2	-0.3	2	-1.8	
1.4	-2.7	-1	-1.2	-2.2	1.3	2.4	0.2	5.1	

Stejným postupem jako v předchozím příkladu byly získány ML odhady parametrů modelu (13) iterativním postupem podle algoritmu uvedeném v Tab. 7. Postup výpočtu a výsledné hodnoty parametrů jsou shrnuty v Tab. 10 a Tab. 11.



Obr. 20 Naměřená data z Hayes-Lemmerz (a) a jejich hustota: nultý odhad (b) a vypočítaný ML-odhad (c)

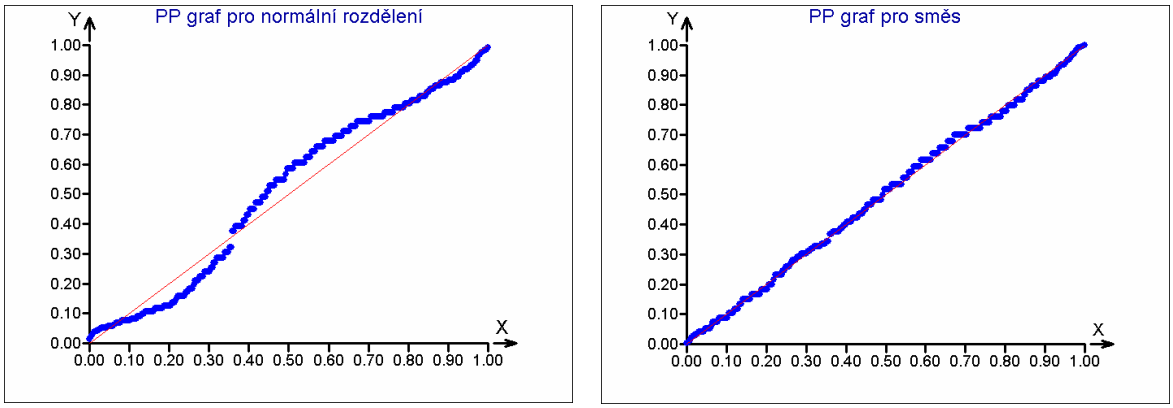
Tab. 10 Tabulka iterativního výpočtu maximálně věrohodného odhadu

Index	m1	log(s1)	m2	log(s2)	atanh(alfa)	-log(L)	||G||	dump	min_eig	N_neg
0	2	0	-2	-0.5	0	556.6001	118.3084	0.1		
1	1.976	0.036	-1.983	-0.475	0.134	546.4	118.308	0.19	34.42	0
2	1.946	0.081	-1.96	-0.441	0.291	536.6	96.803	0.271	31.18	0
3	1.913	0.127	-1.934	-0.4	0.441	529.4	72.089	0.344	27.95	0
4	1.88	0.169	-1.908	-0.359	0.566	525.1	48.982	0.41	24.98	0
5	1.85	0.203	-1.886	-0.324	0.659	523.1	30.365	0.469	22.17	0
6	1.825	0.229	-1.872	-0.3	0.721	522.3	17.111	0.522	19.69	0
7	1.808	0.245	-1.863	-0.285	0.759	522.1	8.696	0.57	17.79	0
8	1.798	0.254	-1.86	-0.278	0.778	522.1	3.948	0.613	16.56	0
9	1.793	0.258	-1.858	-0.275	0.787	522	1.591	0.651	15.89	0
10	1.791	0.26	-1.858	-0.274	0.791	522	0.569	0.686	15.58	0
11	1.79	0.26	-1.858	-0.274	0.792	522	0.181	0.718	15.46	0

12	1.79	0.26	-1.858	-0.273	0.793	522	0.052	0.746	15.42	0
13	1.79	0.26	-1.858	-0.273	0.793	522	0.013	0.771	15.41	0
14	1.79	0.26	-1.858	-0.273	0.793	522	0.003	0.794	15.4	0
15	1.79	0.26	-1.858	-0.273	0.793	522	0.001	0.815	15.4	0
16	1.79	0.26	-1.858	-0.273	0.793	522	0	0.833	15.4	0
17	1.79	0.26	-1.858	-0.273	0.793	522	0	0.85	15.4	0
18	1.79	0.26	-1.858	-0.273	0.793	522	0	0.865	15.4	0
19	1.79	0.26	-1.858	-0.273	0.793	522	0	0.878	15.4	0

Tab. 11 Maximálně věrohodné odhady parametrů směsi 2 normálních rozdělení:

μ_1	σ_1	μ_2	σ_2	N_1	N_2	-log$L$
1.7897	1.2975	-1.858	0.7608	176.9	80.1	522.048



(A) P-P graf pro normální rozdělení **(B) P-P graf pro směs dvou normálních rozdělení**

Obr. 21 Porovnání shody dat s modelem

Podle Tab. 11 přísluší tedy přes 30% dat k rozdělení s nižší střední hodnotou (odhad μ_1=1.79), což by bylo v dobré shodě s předpokládanou třetinou odlišné produkce. Tato metoda (resp. její odhady) je velmi citlivá na přítomnost vybočujících hodnot. V uvedeném výběru byly původně dvě hodnoty 8 a 8.1, které byly rozpoznány jako chybně změřené a odstraněny ještě před analýzou. Pokud by byla tato dvě nepříliš vybočující data ve výběru ponechána, vedlo by to ke snížení odhadu rozsahu N_2 z 80 na 65, a tedy příslušného podílu 100(1 – α) z 31 na pouhých 25%.

4. Detekce a identifikace změny

4.1. Skoková změna střední hodnoty

Modely skokové změny jsou rozsáhle studovány a publikovány od sedmdesátých let dvacátého století jak v klasické statistické literatuře, tak v aplikacích, především z oblasti enviromentálních, biologických a ostatních přírodních věd, např. [17] až [26]. Problém detekce změny (change point detection) je v nejjednodušší podobě definován pomocí modelu

$$y_i = \begin{cases} \mu + \varepsilon_i & i = 1, 2, ..., k \\ \mu + \alpha + \varepsilon_i & i = k+1, ..., n \end{cases} \quad \textbf{(24)}$$

s pararametry μ, α a k, které je třeba odhadnout. Změna střední hodnoty procesu v okamžiku $i = k$; $1 < k < n$ nastala, jestliže $\alpha \neq 0$.

Problém detekce změny je do jisté míry podobný problému z předcházející kapitoly, jenže výběry z jednoho a druhého rozdělení následují po sobě. V technické a výzkumné praxi se s ním setkáváme velmi často, když dochází k náhlé nechtěné fyzikální změně v systému a je třeba tuto změnu detekovat a identifikovat čas, kdy k ní došlo. Nejčastěji používané kritérium pro testování H_0: $\alpha = 0$ proti H_A: $\alpha \neq 0$ je založené na částečných kumulativních součtech

$$T_n = \max_{1 \le k \le n} \left| \sum_{i=1}^{k} \frac{(y_i - \bar{y}_n)}{\sqrt{n s_n^2}} \right| \quad \textbf{(25)}$$

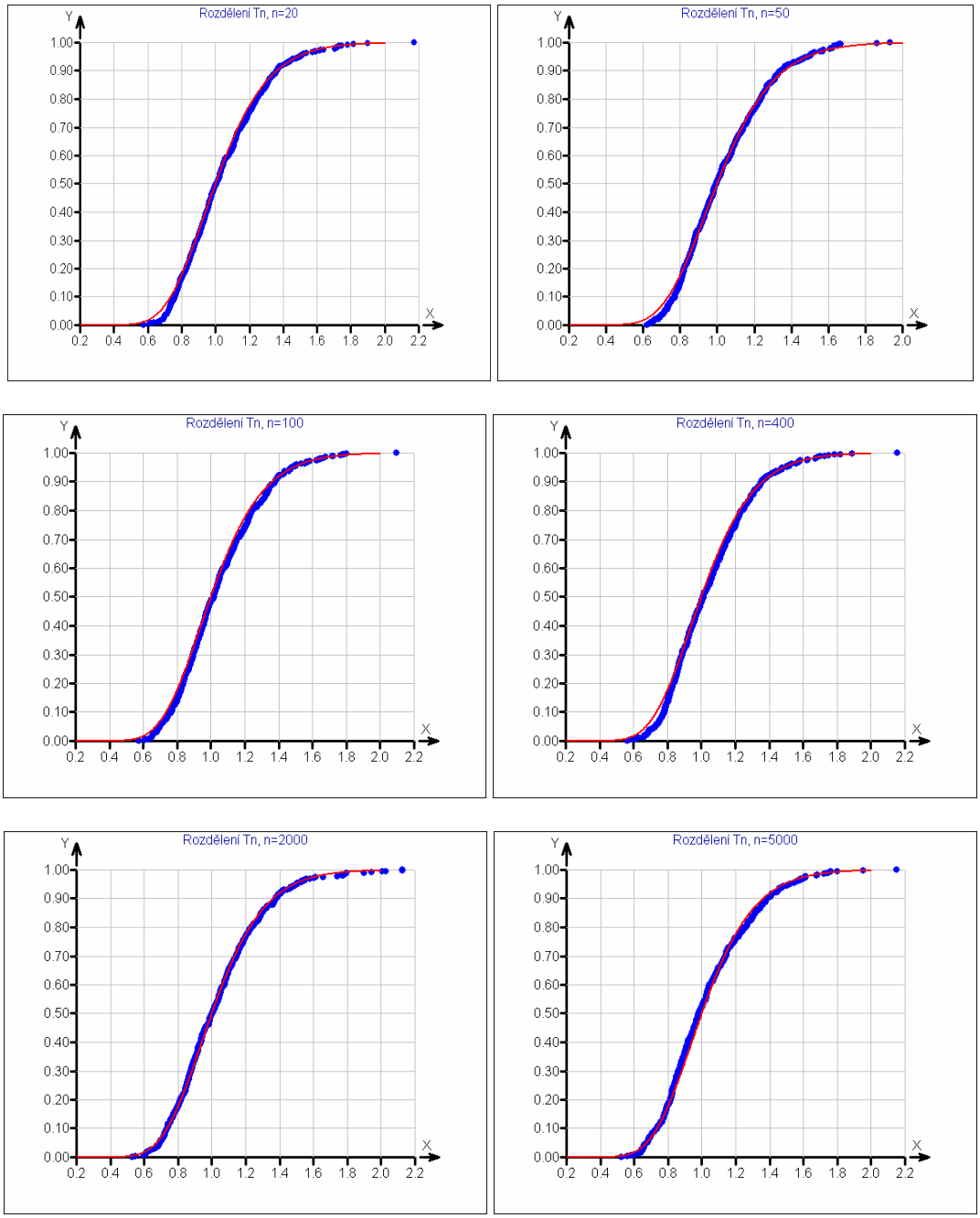
Běžně se uvádí (např. [20]), že statistika T_n má přibližně rozdělení $t(n)$, avšak přesnější v rozmezí $10 < n < 5000$ se na základě simulací ukazuje empirická aproximace Fisherovým rozdělením

$$T_{nF} = T_n + \frac{1}{3} - \frac{1}{50} \ln n \sim F(70, 70), \quad \textbf{(26)}$$

jak je ilustrováno porovnáním simulovaných statistik s distribuční funkcí $F(70, 70)$ pro n=20, 50, 100, 400, 2000 a 5000 na Obr. 22.

Tab. 12 Skript pro simulaci rozdělení statistiky T_{nF}

```
//deletevars
n=100 //Délka řady
n1=50 // Bod zlomu
a=0.0 // Skok
nb=500 // Počet simulací
for(i=1,nb) // Cyklus simulací
{
x=normalr(n)
shift=vec(rep(0,n1),rep(a,n-n1))
x=x+shift
prum=average(x)
smo=sqrt(var(x))
tn1=cusum(x-prum)/(smo*sqrt(n))
tn0[i]= max(abs(tn1))
}
xg=sort(1,tn0)
xg2=xg+1/3-1/50*ln(n)
yg=seq(1/nb/2,1-1/nb/2,count=nb)
xgf=seq(0.2,2,count=1000)
plot(xg2,yg,main="Rozdělení Tn, n="+n)
plotadd(xgf,fisherp(xgf,70,70),type=line,color=3,width=2)
```



Obr. 22 Porovnání simulovaných statistik T_n+1/3-1/50 ln(n) s rozdělením F(70,70) pro různá n

4.2. Skoková změna parametrů spojitého regresního modelu

Jako jisté zobecnění skokové změny střední hodnoty lze chápat skokovou změnu parametrů regresního modelu, kterou lze obecně pro lineární hranici změny zapsat jako

$$Y = \begin{cases} g(\mathbf{x}, \boldsymbol{\theta}_1) & \mathbf{x}^T\mathbf{c} - d < 0 \\ g(\mathbf{x}, \boldsymbol{\theta}_2) & \mathbf{x}^T\mathbf{c} - d \geq 0 \end{cases}, \qquad \textbf{(27)}$$

kde Y je závisle proměnná, $\mathbf{x}$ je vektor nezávislé proměnné, $\boldsymbol{\theta} \neq \boldsymbol{\theta}_1$ jsou vektory regresních parametrů a $\mathbf{c}$ a d jsou lineární podmínky. Tato úloha je bohatě diskutována v literatuře jak ze statistického, tak i z aplikačního hlediska, [82] - [117], případně i v souvislosti se splajny a lokální regresí [118] - [131]. Při známém vektoru $\mathbf{c}$ jde o dvě nezávislé regresní úlohy, avšak je-li $\mathbf{c}$ neznámé a je cílem jej odhadnout, je jednou z možností reparametrizace regresního modelu se zabudováním lineárních podmínek jako dalších parametrů. Při tom lze využít fyzikálně opodstatněného požadavku spojitosti modelu, takže

$$g(\mathbf{x}, \boldsymbol{\theta}_1) = g(\mathbf{x}, \boldsymbol{\theta}_2) \; pro \; \mathbf{x}^T \mathbf{c} = d \,. \tag{28}$$

Dá se ukázat, že reparametrizace při této podmínce a lineárním modelu g nevede ke zvýšení počtu parametrů, avšak reparametrizovaný model g($\mathbf{x}$, $\boldsymbol{\theta}$) již není lineární v parametrech, tedy $\partial g / \partial \theta_i$ je funkcí $\boldsymbol{\theta}$ pro některá i a je nutné hledat odhady parametrů pouze iterativně, nelineární regresí. Někteří autoři poukazují na nespojitost věrohodnostní funkce na hranici zlomu, a teoretickou neodhadnutelnost maximálně věrohodných odhadů avšak pro většinu reálných úloh lze nalézt jediné minimum součtu čtverců a vyhovující asymptotický odhad parametrů $\boldsymbol{\theta}$ a kovarianční matice cov($\boldsymbol{\theta}$).

Nejjednodušší model tohoto typu je regresní přímka se zlomem v bodě $x = c$ (angl. *broken line regression*, nebo *segmented regression*)

$$Y = \begin{cases} \alpha_1 + \alpha_2 x & x < c \\ \beta_1 + \beta_2 x & x \geq c \end{cases}. \tag{29}$$

Z podmínky spojitosti v c plyne model s parametry $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \beta_2, c)$

$$Y = \mathrm{h}(c - x)(\alpha_1 + \alpha_2 x) + \mathrm{h}(c - x)\left[\alpha_1 + c(\alpha_2 - \beta_2) + \beta_2 x\right], \tag{30}$$

kde skoková funkce h(x) má hodnotu 1 je-li argument nezáporný, jinak 0. Minimalizuje-li se součet čtverců $S(\boldsymbol{\theta})$ = ||Y – g($\mathbf{x}$, $\boldsymbol{\theta}$)||2, lze odhadnout kovarianční matici parametrů jako $\mathrm{cov}(\boldsymbol{\theta}) = \sigma^2 \mathbf{H}\left(g(\mathbf{x}, \boldsymbol{\theta})\right), \boldsymbol{\theta} = \boldsymbol{\theta}^*$, kde $\boldsymbol{\theta}^*$ je hodnota parametrů v minimu, $\mathbf{H}$ je matice druhých derivací modelu g($\mathbf{x}$, $\boldsymbol{\theta}$) s prvky $H_{ij} = \partial g(\mathbf{x}, \boldsymbol{\theta}) / \partial \theta_i \partial \theta_j$ a σ^2 je reziduální rozptyl. Odtud lze získat také interval spolehlivosti odhadu polohy zlomu c, který představuje v praxi často zásadní informaci. Dále lze testovat významnost, nebo shodnost parametrů. Účelová funkce $S(\boldsymbol{\theta})$ je obvykle konvexní v dostatečně velkém okolí $\boldsymbol{\theta}^*$ a Gaussova metoda tedy konverguje rychle. Navíc není obvykle obtížné nalézt vhodné počáteční odhady, neboť parametry mají zřejmý geometrický význam. Nezbytný je především správný odhad polohy zlomu c, neboť je-li počáteční odhad c_0 mimo rozsah $\mathbf{x}$, derivační optimalizace $S(\boldsymbol{\theta})$ bude selhávat, protože pak je příslušná parciální derivace nulová,

$$\frac{\partial S(\boldsymbol{\theta})}{\partial c} = 0 \quad pro\ c \in (-\infty, \min x) \cup (\max x, \infty). \tag{31}$$

Model (30) lze zobecňovat jednak do více dimenzí nezávisle proměnné, jednak rozšířit na složitější modely s jednorozměrnou nezávisle proměnnou. Omezíme-li se na jednorozměrnou nezávisle proměnnou, lze jako další příklady uvést dvojitý kvadratický model se šesti parametry

$$Y = \mathrm{h}(c-x)\left(\alpha_1 + \alpha_2 x + \alpha_3 x^2\right) + \mathrm{h}(c-x)\left[\alpha_1 + c(\alpha_2 - \alpha_5) + c^2(\alpha_3 - \alpha_6) + \alpha_5 x + \alpha_6 x^2\right]. \tag{32}$$

Někteří autoři navrhli model s hladkým průběhem v bodu zlomu zavedením hladké přechodové funkce místo Heavisidova skoku. Tyto funkce mohou být například typu $\frac{1}{1+e^x}$, nebo $\frac{1}{2} + \frac{1}{2}\tanh(x)$. Regresní model s přechodovou funkcí má tedy například tvar

$$Y = \frac{1}{1+e^{p(c-x)}}(\alpha_1 + \alpha_2 x) + \frac{1}{1+e^{p(x-c)}}\left[\alpha_1 + c(\alpha_2 - \beta_2) + \beta_2 x\right], \tag{33}$$

kde p je předem zvolená pevná konstanta definující strmost přechodu.

Výhodou této přechodové funkce je odstranění nespojitosti věrohodnostní funkce, nevýhodou je nutnost zavedení nového parametru s významem strmosti přechodové funkce, který se obvykle nedá odhadnout, nebo se odhaduje velmi obtížně. Proto je obvykle nutné jej definovat předem jako pevnou vhodně zvolenou empirickou konstantu. Za předpokladu známého rozptylu má pak podmínka maximální věrohodnosti tvar soustavy (34), jejíž levé strany jsou spojité a hladké pro $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \beta_1, c) \in \mathbb{R}^4$ (s použitím Maple 9)

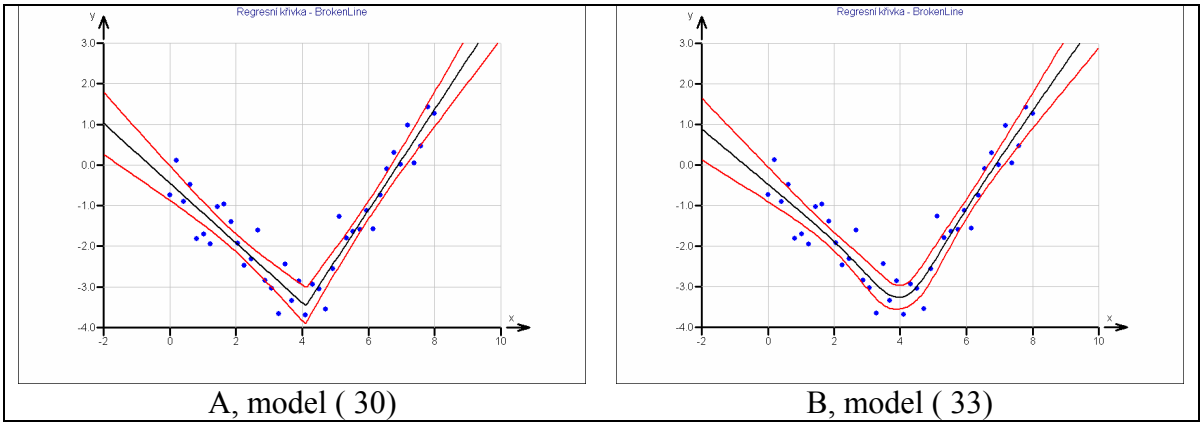
$$\begin{aligned}
\frac{\partial \ln L(\mathbf{x}, y, \boldsymbol{\theta})}{\partial \alpha_1} &= \sum_X 2\left(y - \frac{\alpha 1 + \alpha 2\, X}{1 + \mathbf{e}^{(X-c)}} - \frac{\alpha 1 + c\,(\alpha 2 - \beta 2) + \beta 2\, X}{1 + \mathbf{e}^{(c-X)}}\right)\left(-\frac{1}{1 + \mathbf{e}^{(X-c)}} - \frac{1}{1 + \mathbf{e}^{(c-X)}}\right) &&= 0 \\
\frac{\partial \ln L(\mathbf{x}, y, \boldsymbol{\theta})}{\partial \alpha_2} &= \sum_X 2\left(y - \frac{\alpha 1 + \alpha 2\, X}{1 + \mathbf{e}^{(X-c)}} - \frac{\alpha 1 + c\,(\alpha 2 - \beta 2) + \beta 2\, X}{1 + \mathbf{e}^{(c-X)}}\right)\left(-\frac{X}{1 + \mathbf{e}^{(X-c)}} - \frac{c}{1 + \mathbf{e}^{(c-X)}}\right) &&= 0 \\
\frac{\partial \ln L(\mathbf{x}, y, \boldsymbol{\theta})}{\partial \beta_2} &= \sum_X -\frac{2\left(y - \frac{\alpha 1 + \alpha 2\, X}{1 + \mathbf{e}^{(X-c)}} - \frac{\alpha 1 + c\,(\alpha 2 - \beta 2) + \beta 2\, X}{1 + \mathbf{e}^{(c-X)}}\right)(X - c)}{1 + \mathbf{e}^{(c-X)}} &&= 0 \\
\frac{\partial \ln L(\mathbf{x}, y, \boldsymbol{\theta})}{\partial c} &= \sum_X 2\left(y - \frac{\alpha 1 + \alpha 2\, X}{1 + \mathbf{e}^{(X-c)}} - \frac{\alpha 1 + c\,(\alpha 2 - \beta 2) + \beta 2\, X}{1 + \mathbf{e}^{(c-X)}}\right)\left(-\frac{(\alpha 1 + \alpha 2\, X)\,\mathbf{e}^{(X-c)}}{(1 + \mathbf{e}^{(X-c)})^2} - \frac{\alpha 2 - \beta 2}{1 + \mathbf{e}^{(c-X)}} + \frac{(\alpha 1 + c\,(\alpha 2 - \beta 2) + \beta 2\, X)\,\mathbf{e}^{(c-X)}}{(1 + \mathbf{e}^{(c-X)})^2}\right) &&= 0.
\end{aligned} \tag{34}$$

Dalším výhodným důsledkem zavedení přechodové funkce je snížení vlivu měření v okolí zlomu, vyjádřené například diagonálními prvky H_{ii} matice $\mathbf{H}=\mathbf{J}(\mathbf{J}^T\mathbf{J})^{-1}\mathbf{J}^T$, kde jakobián $J_{ij}=\frac{\partial g(\boldsymbol{\alpha},x_i)}{\partial \alpha_i}$ je parciální derivace modelu v *i*-tém bodě podle *j*-tého parametru.

Možnost využití popsaných modelů ilustrují následující příklady. V prvním příkladě je porovnán regresní model (30) a (33) na simulovaných datech s parametry α_1= -1, α_2=-0.5, β_2=1 a c=4. Z výsledků i grafů je patrné, že zavedení přechodové funkce nemělo významný vliv na bodové ani intervalové odhady parametrů. Skript v jazyce DARWin použitý pro generování dat je uveden v Tab. 13. Pro nelineární regresi byly použity derivační metody programu QCExpert.

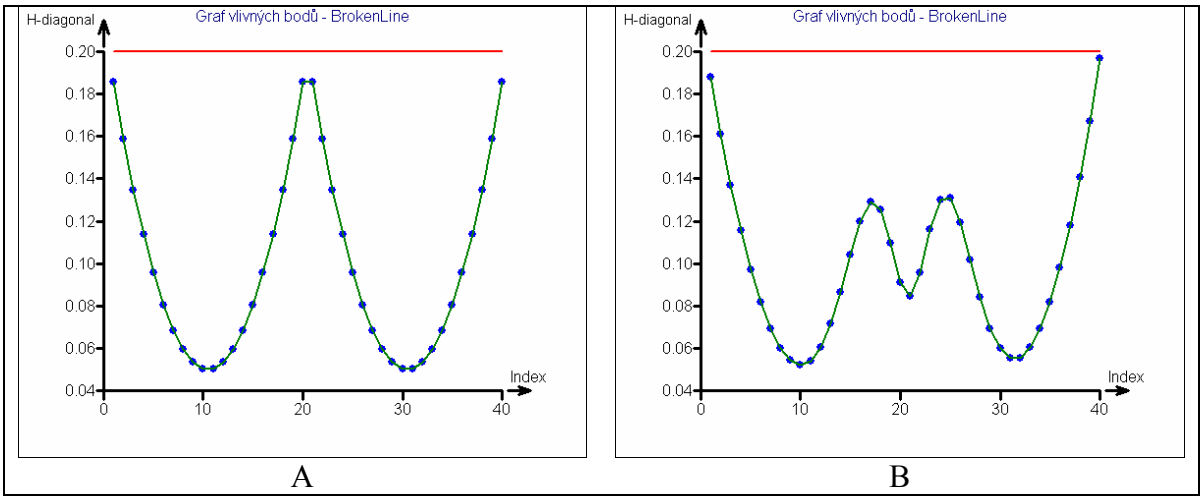
Tab. 13 Skript pro generování dat se zlomem

```
// Simulace dat pro regresi se zlomem regresi
n=40 //Počet dat
sd=0.5 //Směrodatná odchylka
c=4 // Poloha zlomu
a1=-1;a2=-0.5;b2=1 // Parmetry pro model s přímkovými segmenty
//a1=-1;a2=0.5;a3=-0.5;a5=-3;a6=0.4 // Parametry pro model
s parabolickými segmenty
x=seq(0,8,count=n)
// Model s přímkovými segmenty:
y=heav(c-x)*(a1+a2*x)+heav(x-c)*(a1+c*(a2-b2)+b2*x)
// Model s parabolickými segmenty:
//y=heav(c-x)*(a1+a2*x+a3*x^2)+heav(x-c)*(a1+c*(a2-a5)+c^2*(a3-
a6)+a5*x+a6*x^2)
y=y+normalr(n,mean=0,sdev=sd)
plot(x,y)
data=bind(x,y)
copy(data,"BrokenLine")
```



A, model (30) B, model (33)

Obr. 23 Model s přímkovými větvemi bez (A) a se (B) spojitou přechodovou funkcí



Obr. 24 Vliv jednotlivých dat v modelech z Obr. 23 vyjádřených diagonálními prvky H_{ii}

Tab. 14 Výsledky regrese k modelu (30), Obr. 23 A

Název úlohy :	BrokenLine
Data:	Všechna
Nezávisle proměnné :	X
Závisle proměnná :	Y
Hladina významnosti :	0.05
Počet stupňů volnosti :	36
Kvantil t(1-alfa/2,n-p) :	2.028094001
Kvantil F(1-alfa,m,p-m) :	10.12796449
Metoda :	Nejmenší čtverce
Počet platných řádků :	40
Počet parametrů :	4
Metoda optimalizace :	Gradient s pevným krokem
Model :	[Y] ~ (p1+p2*[X])*lt([X],p4)+(p1+p4*(p2-p3)+p3*[X])*ge([X],p4)
Počáteční hodnoty parametrů :	
P1	0
P2	-1
P3	1
P4	5
Výpočet	
Počet iterací :	19
Ukončení výpočtu :	Konvergence
Doba výpočtu :	0.42 s
Max. počet iterací :	999999
Terminační kritérium :	1E-008

Odhady parametrů	Parametr	Směr. odchylka	Dolní mez	Horní mez
P1	-0.4454268492	0.2091025303	-0.8695064365	-0.02134726186
P2	-0.7359560965	0.09172768424	-0.9219884627	-0.5499237304
P3	1.241137896	0.09171415655	1.055132965	1.427142826
P4	4.102575634	0.1555236949	3.787158961	4.417992306

Korelační matice parametrů :		P1	P2	P3	P4
	P1	1	-0.8548504143	-5.556549903E-005	-0.3661129531
	P2	-0.8548504143	1	4.750020673E-005	0.6424752908
	P3	-5.556549903E-005	4.750020673E-005	1	0.5815238919
	P4	-0.3661129531	0.6424752908	0.5815238919	1

Tab. 15 Výsledky regrese k modelu (33), Obr. 23 B

Název úlohy :	BrokenLine
Data:	Všechna
Nezávisle proměnné :	X
Závisle proměnná :	Y
Hladina významnosti :	0.05
Počet stupňů volnosti :	36
Kvantil t(1-alfa/2,n-p) :	2.028094001
Kvantil F(1-alfa,m,p-m) :	10.12796449
Metoda :	Nejmenší čtverce
Počet platných řádků :	40
Počet parametrů :	4
Metoda optimalizace :	Gauss-Newton
Model :	[Y] ~ (p1+p2*[X])*(1/(1+exp(2*([X]-p4))))+(p1+p4*(p2-p3)+p3*[X])*(1/(1+exp(2*(p4-[X]))))
Počáteční hodnoty parametrů :	
P1	0
P2	-1
P3	1
P4	5
Výpočet	
Počet iterací :	21
Ukončení výpočtu :	Konvergence
Doba výpočtu :	0.23 s
Max. počet iterací :	999999
Terminační kritérium :	1E-008

Odhady parametrů	Parametr	Směr. odchylka	Dolní mez	Horní mez
P1	-0.4717137835	0.2104409117	-0.898507734	-0.04491983304
P2	-0.6791832815	0.089154221	-0.8599964222	-0.4983701407
P3	1.171625809	0.09522258521	0.9785054548	1.364746162
P4	4.087022165	0.1734106544	3.735329057	4.438715273

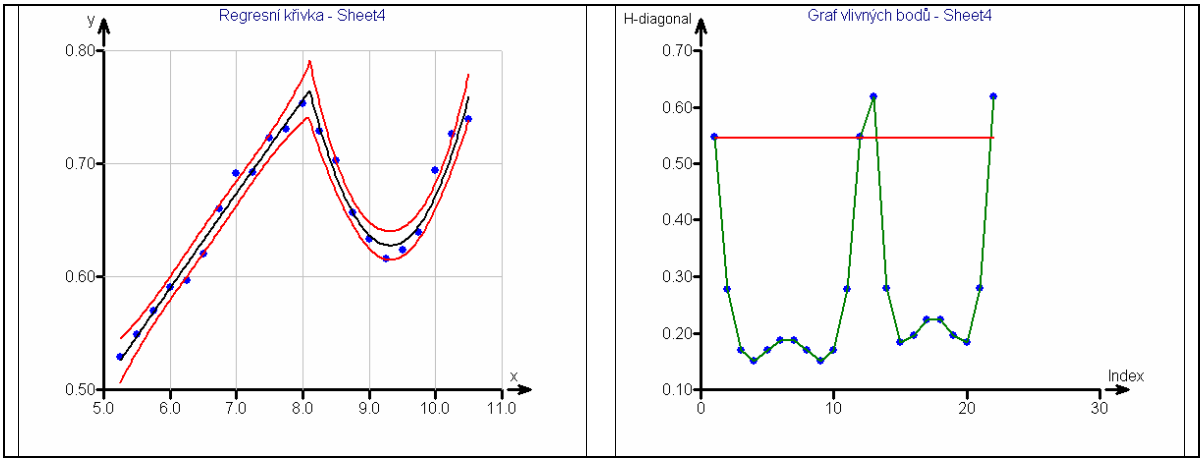
Korelační matice parametrů :		P1	P2	P3	P4
	P1	1	-0.8639114112	-0.09823391684	-0.397362888
	P2	-0.8639114112	1	0.1375130212	0.6614004508
	P3	-0.09823391684	0.1375130212	1	0.6733405645
	P4	-0.397362888	0.6614004508	0.6733405645	1

Použití modelu (32) je ilustrováno na vývoji čtvrtletního exportu zahraničního obchodu v letech 2Q 2005 až 3Q 2010. Data, viz Tab. 16, pocházejí ze zdroje Českého statistického úřadu (http://www.czso.cz/csu/csu.nsf/i/tab_vs/$File/tab_vs_4q10.xls). Data vykazují náhlý pokles na začátku roku 2008 v důsledku očekávání ekonomické recese, v průběhu dalších 2 let pak postupný progresivní nárůst. I na tomto příkladu je ilustrována použitelnost bezparametrické přechodové funkce. Model (32) bez přechodové funkce (na Obr. 25) i s přechodovou funkcí (na Obr. 26) poskytuje prakticky shodné výsledky, včetně odhadu polohy zlomu. Za povšimnutí stojí výrazné snížení vlivu dat v okolí zlomu v případě použití přechodové funkce. Shrnutí výsledků regrese je v Tab. 17 pro model s ostrým zlomem a Tab. 18 pro stejný model s přechodovou funkcí. 95% interval spolehlivosti zlomu je (převedeno na datum) asi 27.12.2007 až 17.3.2008.

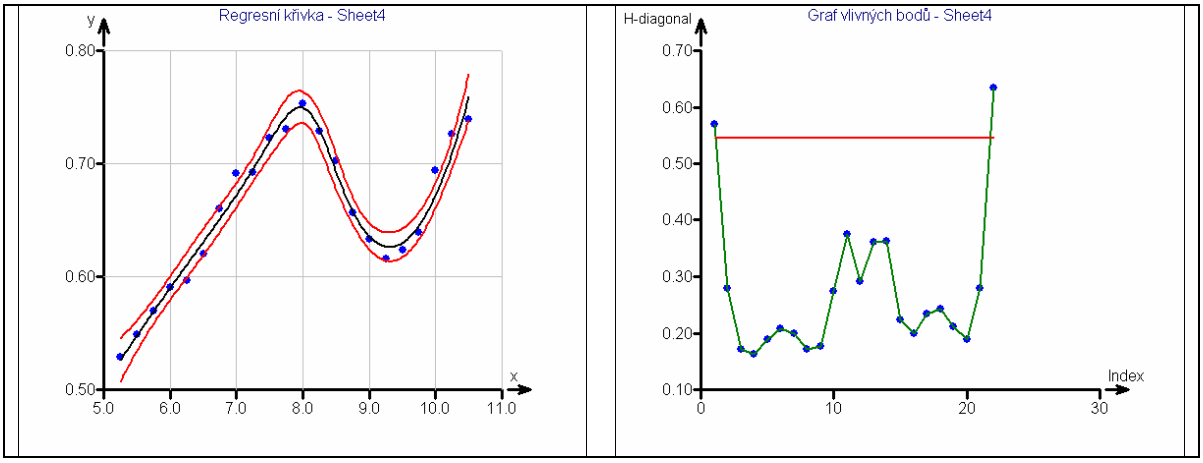
Tab. 16 Export zahraničního obchodu v ČR, 1000 mld. Kč, (rok - 2000)

Rok	5.25	5.5	5.75	6	6.25	6.5	6.75	7	7.25	7.5	7.75
Export	0.528771	0.548745	0.569523	0.590782	0.596908	0.619653	0.660228	0.691069	0.691842	0.722462	0.730593

Rok	8	8.25	8.5	8.75	9	9.25	9.5	9.75	10	10.25	10.5
Export	0.752975	0.729121	0.702245	0.656231	0.633234	0.615668	0.623512	0.639372	0.6942	0.726411	0.739108



Obr. 25 Graf funkce a vlivu jednotlivých bodů pro data z Tab. 16, model (32)



Obr. 26 Graf funkce a vlivu jednotlivých bodů pro data z Tab. 16, model (32) s přechodovou funkcí

Tab. 17 Výsledky regrese k polynomickému modelu bez přechodové funkce, Obr. 25

Nezávisle proměnné :	X
Závisle proměnná :	Y
Hladina významnosti :	0.05
Počet stupňů volnosti :	16
Kvantil t(1-alfa/2,n-p) :	2.119905299
Kvantil F(1-alfa,m,p-m) :	6.607890974
Metoda :	Nejmenší čtverce
Počet platných řádků :	22
Počet parametrů :	6
Metoda optimalizace :	Gauss-Newton

Model :	[Y] ~ (p1+p2*[X]+p3*[X]^2)*lt([X],p4)+(p1+p4*(p2-p5)+p4^2*(p3-p6)+p5*[X]+p6*[X]^2)*ge([X],p4)
Počáteční hodnoty parametrů :	
P1	0
P2	0.1
P3	0
P4	8
P5	-1.5
P6	0.1
Výpočet	
Počet iterací :	20
Ukončení výpočtu :	Konvergence
Doba výpočtu :	0.23 s
Max. počet iterací :	100
Terminační kritérium :	1E-008

Odhady parametrů	Parametr	Směr. odchylka	Dolní mez	Horní mez
P1	0.05476	0.23508	-0.44358	0.55311
P2	0.09349	0.07181	-0.05875	0.24572
P3	-0.00072	0.00541	-0.01219	0.01075
P4	8.10063	0.0527	7.98891	8.21235
P5	-1.74043	0.16139	-2.08256	-1.39831
P6	0.09344	0.0086	0.0752	0.11167

Korelační matice parametrů :		P1	P2	P3	P4	P5	P6
	P1	1.0	-0.99814	0.99308	-0.43268	0.00002	-1.941E-007
	P2	-0.99814	1.0	-0.99834	0.45395	-0.00002	1.9531E-007
	P3	0.99308	-0.99834	1.0	-0.47632	0.00002	-1.958E-007
	P4	-0.43268	0.45395	-0.47632	1.0	-0.60118	0.58712
	P5	0.00002	-0.00002	0.00002	-0.60118	1.0	-0.99943
	P6	-1.9411E-007	1.953E-007	-1.9581E-007	0.58712	-0.99943	1.0

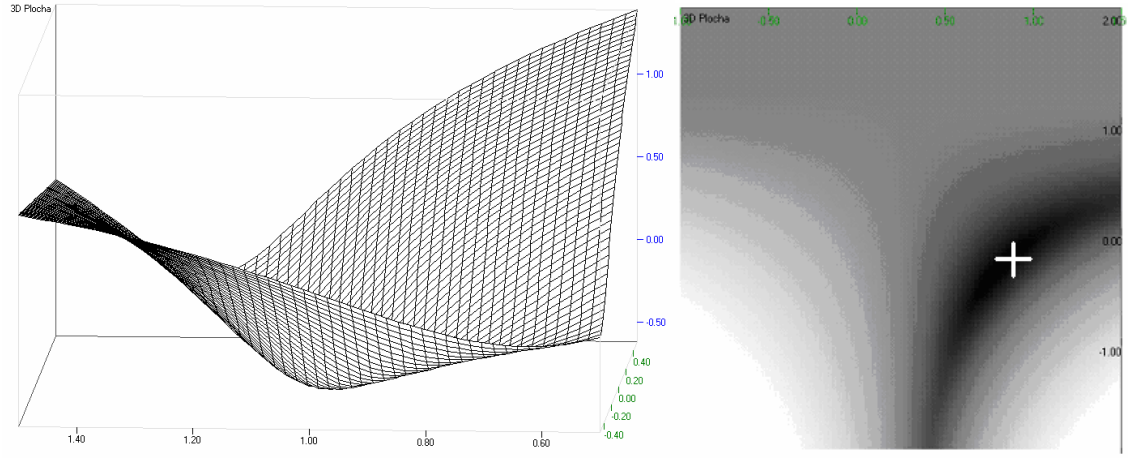
Tab. 18 Výsledky regrese k polynomickému modelu s přechodovou funkcí, Obr. 26

Nezávisle proměnné :	X
Závisle proměnná :	Y
Hladina významnosti :	0.05
Počet stupňů volnosti :	16.0
Kvantil t(1-alfa/2,n-p) :	2.11991
Kvantil F(1-alfa,m,p-m) :	6.60789
Metoda :	Nejmenší čtverce
Počet platných řádků :	22.0
Počet parametrů :	6.0
Metoda optimalizace :	Levenberg-Marquardt
Model :	[Y] ~ (p1+p2*[X]+p3*[X]^2)*(1/(1+exp(4*([X]-p4)))) + (p1+p4*(p2-p5)+p4^2*(p3-p6)+p5*[X]+p6*[X]^2)*(1/(1+exp(4*(p4-[X]))))
Počáteční hodnoty parametrů :	
P1	0.53
P2	-0.04662
P3	0.00881
P4	8.10761
P5	-1.79985
P6	0.09738
Výpočet	
Počet iterací :	75.0
Ukončení výpočtu :	Konvergence
Doba výpočtu :	2.54 s
Max. počet iterací :	100.0
Terminační kritérium :	1E-008

Odhady parametrů	Parametr	Směr. odchylka	Dolní mez	Horní mez
P1	-0.04596	0.24997	-0.57588	0.48396
P2	0.12992	0.07659	-0.03244	0.29228
P3	-0.004	0.00578	-0.01625	0.00825
P4	8.10909	0.07187	7.95673	8.26144
P5	-1.65064	0.168	-2.00677	-1.2945
P6	0.08902	0.00895	0.07005	0.10799

Korelační matice parametrů :		P1	P2	P3	P4	P5	P6
	P1	1.0	-0.9984	0.99412	-0.60873	0.23877	-0.23258
	P2	-0.9984	1.0	-0.99861	0.63121	-0.2485	0.24219
	P3	0.99412	-0.99861	1.0	-0.65393	0.25851	-0.2521
	P4	-0.60873	0.63121	-0.65393	1.0	-0.69789	0.6835
	P5	0.23877	-0.2485	0.25851	-0.69789	1.0	-0.99948
	P6	-0.23258	0.24219	-0.2521	0.6835	-0.99948	1.0

Na Obr. 27 je pro ilustraci zobrazena funkce součtu čtverců $S(\mathbf{x}, \boldsymbol{\theta})$ pro jednoduchý dvouparametrický model se zlomem $Y = \mathrm{h}(x - c)\,\alpha x$. Na této kriteriální ploše, zobrazené jako 3d plocha a jako stínovaný kolmý průmět, je zřejmá spojitost, nelinearita a nekonvexnost součtu čtverců.



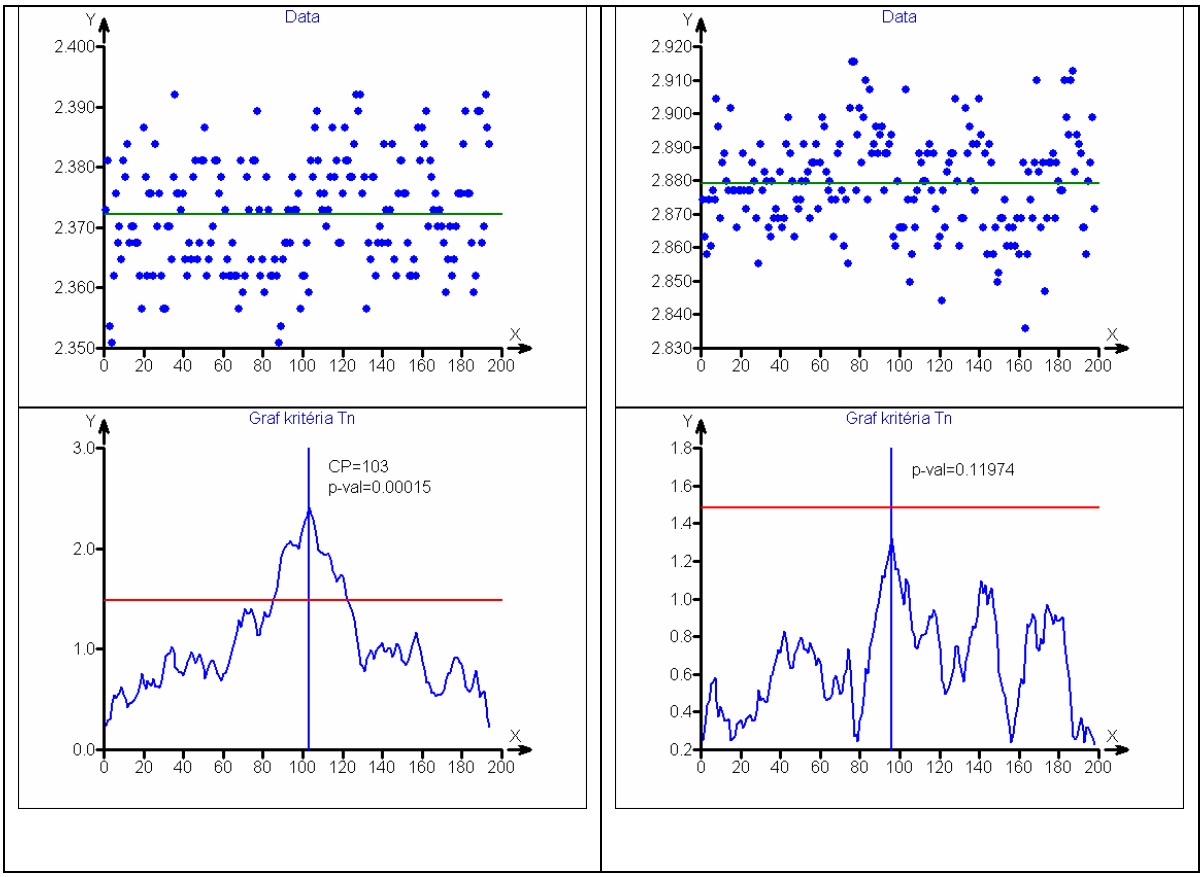
Obr. 27 Tvar kriteriální funkce S(x,θ) pro dvouparametrický model se zlomem

4.3. Aplikace

Data v následující aplikaci pocházejí z kontroly chemického složení sekundárního energetického média v jaderné elektrárně Dukovany. Po provedené změně v technologii a režimu provozu bylo třeba ověřit, které ze čtyř vybraných proměnných (jedná se o vodivost a řízené obsahy solí) na tuto změnu zareagovaly. S použitím modelu (24), kumulativních součtů (25) a kritéria (26) na hladině významnosti α=0.05 se prokázala změna sledované proměnné 1 a 3, změna v proměnných 2 a 4 nebyla prokázána. Grafické výstupy jsou uvedeny na

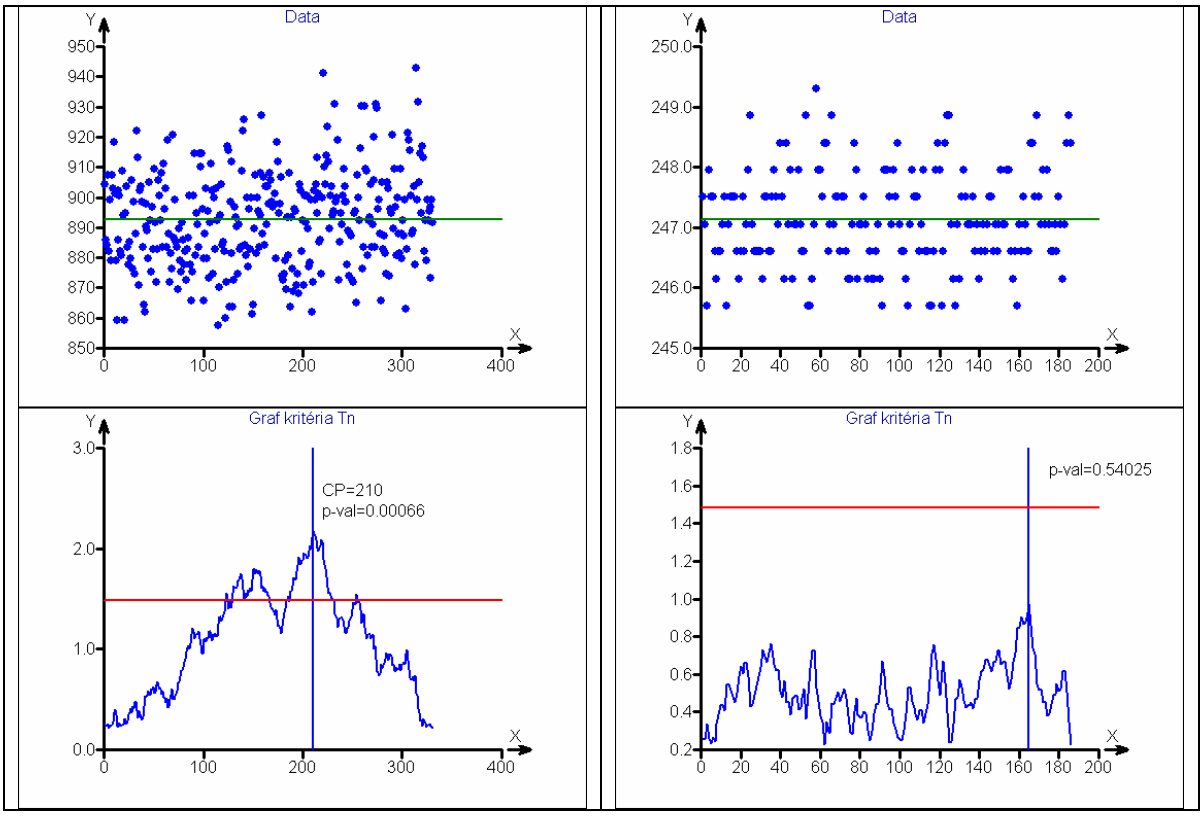
Obr. 28 až

Obr. 31, příslušný skript je v Tab. 19.



Obr. 28 Proměnná 1, změna prokázána, $p=10^{-4}$

Obr. 29 Proměnná 2, změna neprokázána



Obr. 30 Proměnná 3, změna prokázána, $p<10^{-3}$

Obr. 31 Proměnná 4, změna neprokázána

Tab. 19 Skript v jazyce DARWin použitý při testování proměnných

```
// V proměnné x je vstupní časová řada
n=count(x) //Počet dat
prum=average(x) //Průměr
smo=sqrt(var(x))
tn1=cusum(x-prum)/(smo*sqrt(n))
tn2=abs(tn1)+1/3-1/50*ln(n)
tn2max=max(tn2)
fkrit=fisherq(0.95,70,70)
pval=1-fisherp(tn2max,70,70)
ii=eq(tn2,tn2max)
ii2=1:n
imax=ii2[[ii]]

plot(x,main="Data")
lineadd(h=prum,width=2,color=1)
plot(tn2,type=line,width=2, main="Graf kritéria Tn")
lineadd(h=fkrit,color=3,width=2)
lineadd(v=imax,width=2)
if(lt(pval,0.05)){plottextadd(imax+10, max(vec(tn2max,fkrit))+0.4,
"CP="+imax, align=right, textsize=3)}
plottextadd(imax+10, max(vec(tn2max,fkrit))+0.2, "p-val="+round(pval,5),
align=right, textsize=3)
```

Druhá aplikace se týká tří úloh nalezení bodu zlomu v přímkovém modelu (30), strana 41. Jedná se o technologická data (teploty ve stupních celsia měřené v tříminutových intervalech) ze syntetické výroby butadien-styrénových kopolymerů ve středočeském podniku, suroviny pro řadu plastových výrobků. Zlomy v trendech jsou zde kritické momenty a čas bodu zlomu je často rozhodující informace k posouzení probíhajících procesů z hlediska kvality, ekonomiky a bezpečnosti. Pro velký rozsah výběrů zde zdrojová data nejsou uvedena.

První úloha se týká určení času výpadku chlazení reaktoru při exotermické reakci, který vedl v důsledku zvýšení teploty k závažnému technologickému stavu s bezpečnostním rizikem a ekonomickou ztrátou. V bodě $x = 300$ se přitom měnily směny a čas výpadku byl tedy rozhodující pro určení odpovědnosti za škodu.

Byl použit model bez směrnice prvního segmentu (ta je nulová)

$$Y = \mathrm{h}(c-x)(\alpha_1) + \mathrm{h}(c-x)[\alpha_1 + -\beta_2 c + \beta_2 x],$$

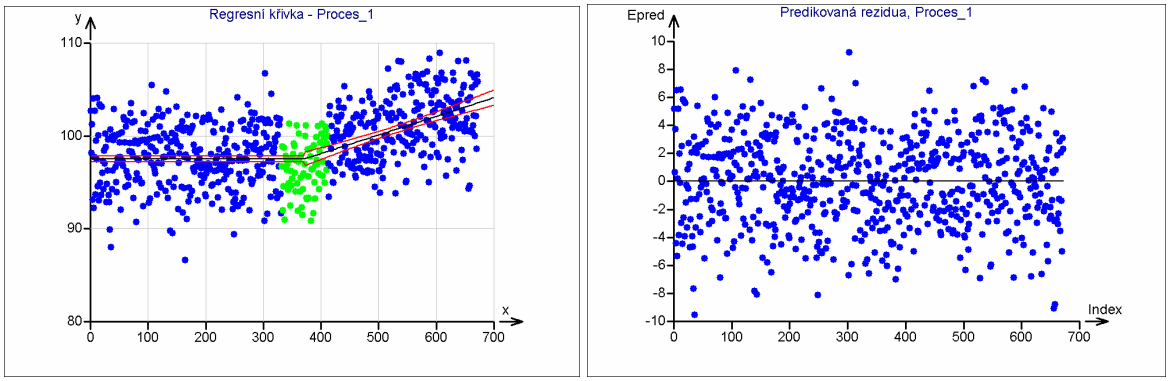
Parametry α_1, β_2 a c jsou kódovány jako p1, p2 a p3. Výsledky shrnuje Tab. 20 a Obr. 32.

Model: [Y] ~ p1*lt([X],p2)+(p1-p3*p2+p3*[X])*ge([X],p2)

Tab. 20 Odhady parametrů regresního modelu

Odhady parametrů	Parametr	Směr. odchylka	Dolní mez	Horní mez
P1	97.5369	0.16679	97.20941	97.86439
P2	371.68736	20.27466	331.87773	411.49699
P3	0.02009	0.00212	0.01592	0.02426

Korelační matice parametrů :		P1	P2	P3
	P1	1.0	0.41154	0.00086
	P2	0.41154	1.0	0.78989
	P3	0.00086	0.78989	1.0



Obr. 32 Regresní model s vyznačeným intervalem spolehlivosti zlomu (zeleně) a predikovaná rezidua

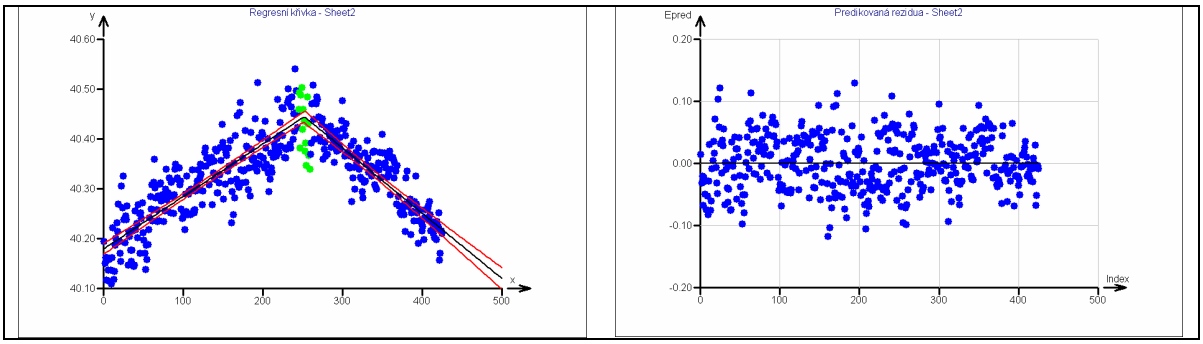
Interval spolehlivosti parametru c = p2 (332, 411) nepodporuje hypotézu, že ke zlomu došlo v bodě $x < 300$ a odpovědnost za havárii tedy leží na druhé směně.

Další dvě ilustrační úlohy 1 a 2 určují bod zlomu z důvodu posouzení kvality procesu a látkových a energetických bilancí. Bod zlomu není možné určit jiným způsobem, než statisticky, neboť nesouhlasí s časem technologických zásahů z důvodů neznámého transportního zpozdění, probíhajících chemických reakcí, turbulencí, apod. V obou případech je použit model (30). Odhady parametrů jsou uvedeny v Tab. 21 a Tab. 22. Interval spolehlivosti polohy zlomu je v grafech na Obr. 33 a Obr. 34 vyznačen zeleně.

Tab. 21 Bodové a intervalové odhady regresních parametů (1)

Odhady parametrů	Parametr	Směr. odchylka	Dolní mez	Horní mez
P1	40.1791	0.00548	40.16833	40.18987
P2	0.00105	0.00004	0.00098	0.00113
P3	-0.00131	0.00007	-0.00144	-0.00118
P4	251.99304	3.61462	244.8881	259.09799

Korelační matice parametrů :		P1	P2	P3	P4
	P1	1.0	-0.86689	-0.00042	0.32171
	P2	-0.86689	1.0	0.00036	-0.55517
	P3	-0.00042	0.00036	1.0	-0.66453
	P4	0.32171	-0.55517	-0.66453	1.0

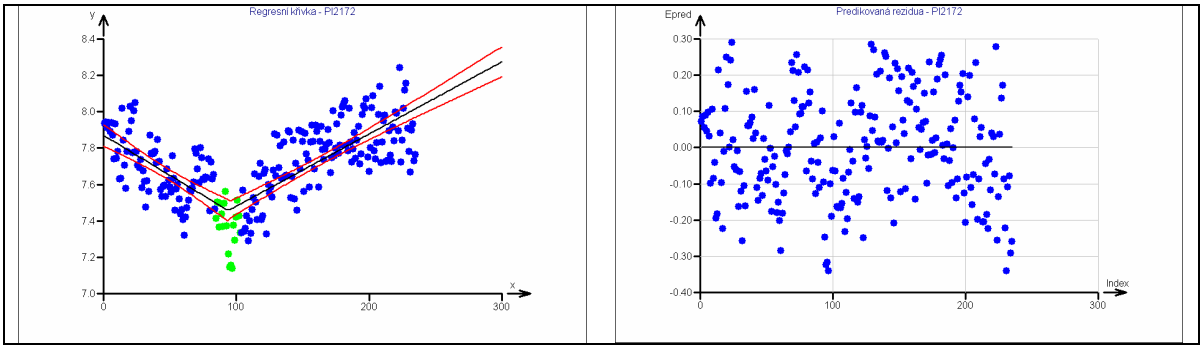


Obr. 33 Regresní funkce a rezidua pro úlohu 1

Tab. 22 Bodové a intervalové odhady regresních parametrů (2)

Odhady parametrů	Parametr	Směr. odchylka	Dolní mez	Horní mez
P1	7.87081	0.03001	7.81168	7.92995
P2	-0.0044	0.00055	-0.0055	-0.00331
P3	0.00397	0.00029	0.00339	0.00455
P4	94.00106	4.58752	84.96233	103.03979

Korelační matice parametrů :		P1	P2	P3	P4
	P1	1.0	-0.86834	-0.00051	-0.39654
	P2	-0.86834	1.0	0.00044	0.67809
	P3	-0.00051	0.00044	1.0	0.54004
	P4	-0.39654	0.67809	0.54004	1.0



Obr. 34 Regresní funkce a rezidua pro úlohu 2

5. Robustní regresní metody, M-odhady

Regresní analýza je jedním z nejpoužívanějších nástrojů v technických aplikacích, např. [132] - [138] se zhruba třemi až čtyřmi nejčastějšími cíli, které se ovšem částečně překrývají. (1) Odhalení vlivu vytypovaných nezávisle proměnných (prediktorů) **x** na zvolenou odezvu *y*, posouzení statistické významnosti vlivu **x** na **y**, případně s cílem proměnnou **y** cíleně ovlivňovat změnami **x**. (2) Odhadnutí hodnot neznámých fyzikálně smysluplných regresních parametrů předem známého modelu. (3) Diagnostika procesů na základě nalezeného modelu, identifikace vybočujících hodnot, nebo změny parametrů. (4) Predikce **y** pro nové hodnoty **x**, případně predikce **x** ze známé hodnoty *y* (kalibrační model).

Obecněji lze rozlišit (nejen v regresi) dva přístupy, viz též [135].

První, empirický a typický spíše pro obory mimo klasickou statistiku, chápe přírodní fyzikální mechanismy jako černou skříňku (black box) a snaží se nalézt vztahy mezi vstupním vektorem nezávisle poměnné **x** a výstupním vektorem **y** reakce studovaného fyzikálního (sociologického, ekonomického, atd.) systému, či fenoménu pomocí empirických stochastických algoritmů. Jako zástupce prvního přístupu lze jmenovat například regresní a klasifikační stromy, neuronové sítě, PLS (partial least squares), SVM (support vector machines), neuronové časové řady.

Druhý, modelový přístup, považuje výstupní vektor **y** za manifestaci (náhodný výběr) známého statistického modelu, jehož neznámé parametry se snaží odhadnout. Ze zástupců druhého přístupu jmenujme třeba lineární a nelineární regresní modely, logistickou regresi, Coxovy modely, modely ARIMA, Fourierovu analýzu. Jelikož oba přístupy jsou z hlediska aplikací přínosné, dotkneme se prvního (v kapitole 6) i druhého (v kapitole 5) a pokusíme se na realizovaných projektech naznačit jejich použití.

Tato kapitola navazuje na zmínku o robustních odhadech v odstavcích 3.1 a 3.2.

V následujících odstavcích se budeme zabývat některými robustními postupy lineární regrese, jejich vlastnostmi a technologickou aplikací. Ačkoliv jsou robustní regresní metody publikovány daleko více [139], [183], než postupy klasické a o jejich výhodách není pochyb, k jejich širšímu povědomí na technických a výzkumných pracovištích a používání v praxi zatím nedochází, i když v ČR existuje tradiční vynikající robustní škola a dokonce i každoroční silně obsazená konference ROBUST pořádaná JČMF. Z mnoha robustních postupů (L_p-odhady, M-odhady, R-odhady, L-odhady, LTS, LMS, atd.) se budeme věnovat aplikaci poměrně spolehlivých M-odhadů , které mají dobře prozkoumanou asymptotickou teorii, numericky jsou relativně snadno implementovatelné a intuitivně pochopitelné pro nestatistiky. Dále se zmíníme o některých praktických vlastnostech L_1-odhadů.

5.1. M-odhady

Jak bylo zmíněno v odst. 3.2, M-odhady minimalizují součet

$$S_M(\boldsymbol{\theta}) = \sum_{i=1}^{n} \rho\left(y_i - g(\mathbf{x})\right) = \sum_{i=1}^{n} \rho(e_i), \qquad (35)$$

který lze pro konkrétní funkci ρ převést na váženou metodu nejmenších čtverců. V případě lineárního regresního modelu lze regresní koeficienty $\boldsymbol{\alpha}$ odhadnout pomocí vztahu

$$\mathbf{a} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}\,, \qquad (36)$$

kde **a** je odhad **α** a **W** je diagonální váhová matice s prvky $w(e_i)$, viz (10), strana 19. Pro získání váhových koeficientů w_{ii} je třeba postup (36) iterativně opakovat s vhodným terminačním kritériem. Při iterativním postupu se obvykle vychází z jednotkové matice $\mathbf{W}_0=\mathbf{I}$, která odpovídá obyčejné metodě nejmenších čtverců (MNČ). První odhady parametrů MNČ pak v následujících iteracích konvergují k M-odhadům v minimu (35), iterační proces je pak ukončen vhodně zvolenou terminační podmínkou, například $\|\mathbf{a}_i - \mathbf{a}_{i-1}\| < 10^{-8}$.

$$\begin{aligned}
&\mathbf{W}_0 := \mathbf{I}\\
&i := 0\\
&\textit{repeat}\\
&\mathbf{a}_i := \left(\mathbf{X}^T\mathbf{W}_i\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}_i\mathbf{y}\\
&\mathbf{e}_i := \mathbf{y} - \mathbf{X}\mathbf{a}_i\\
&\mathbf{W}_{i+1} := \mathrm{diag}\left(\mathrm{w}(\mathbf{e}_i)\right)\\
&i := i + 1\\
&\textit{until}\ \mathrm{term}
\end{aligned} \qquad (37)$$

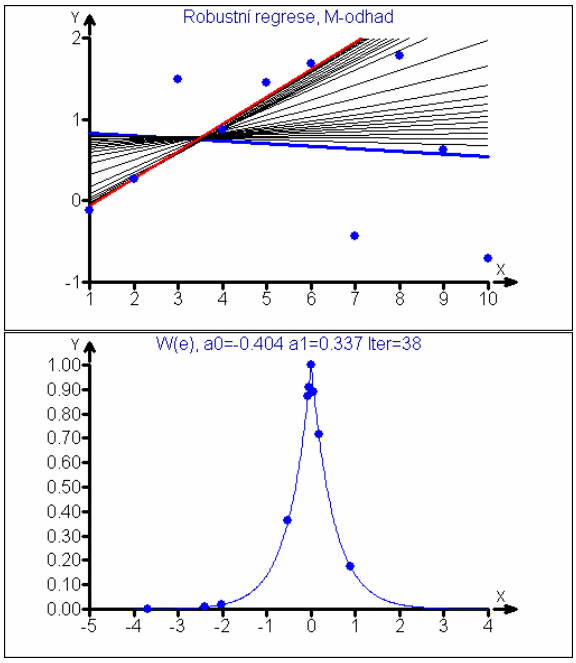
Tento iterační proces, někdy nazývaný IRWLS (*Iteratively Re-Weighted Least Squares*), je ilustrován následujícím příkladem se šesti různými váhovými funkcemi. V příkladu je použit přímkový regresní model $y = \alpha_0 + \alpha_1 x$ s daty uvedenými v tabulce Tab. 23. Skriptem v Tab. 24 jsou vygenerovány grafy na Obr. 35 až Obr. 41. pro vybraných šest často používaných váhových funkcí, resp. odpovídajících funkcí ρ, nebo ψ, viz vztah (10), str. 19. Vějíře přímek zde představují jednotlivé iterace algoritmu (37) od prvního odhadu MNČ (modrá přímka) až po konečný robustní M-odhad (červená přímka). Grafy navíc ukazují výsledný průběh váhové funkce w(e), kde e je vzdálenost y od modelu, s vyznačenými polohami reziduí. Je zřejmé, že během minimalizace (35) došlo až ke změně znaménka směrnice a_1 z –0.03 na 0.3 a absolutního členu a_0 z 0.8 na -0.4. U váhy typu w=1/|e|, Obr. 39 se však výsledné odhady příliš neliší od počátečních odhadů MNČ, což je zřejmě způsobeno nevhodnou volbou konstant váhové funkce, jak bude diskutováno dále.

Tab. 23 Data pro robustní regresi

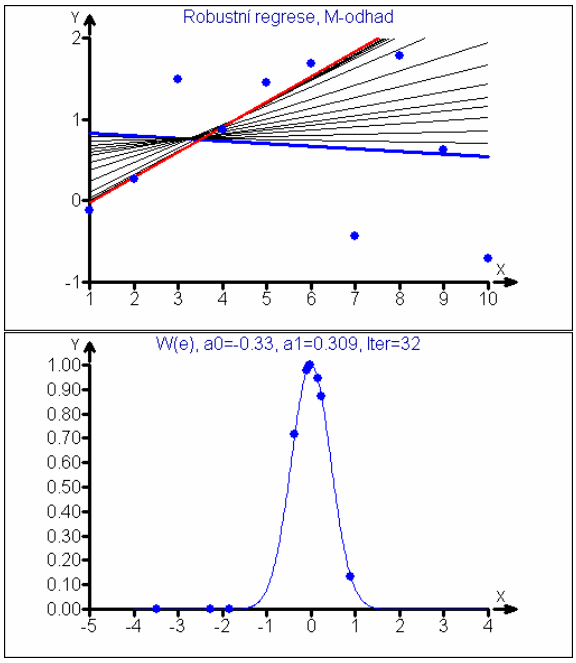
x	1	2	3	4	5	6	7	8	9	10
y	-0.117	0.271	1.497	0.875	1.454	1.680	-0.439	1.780	0.617	-0.717

Tab. 24 Skript v jazyce DARWin pro výpočet M-odhadů parametrů regresního modelu

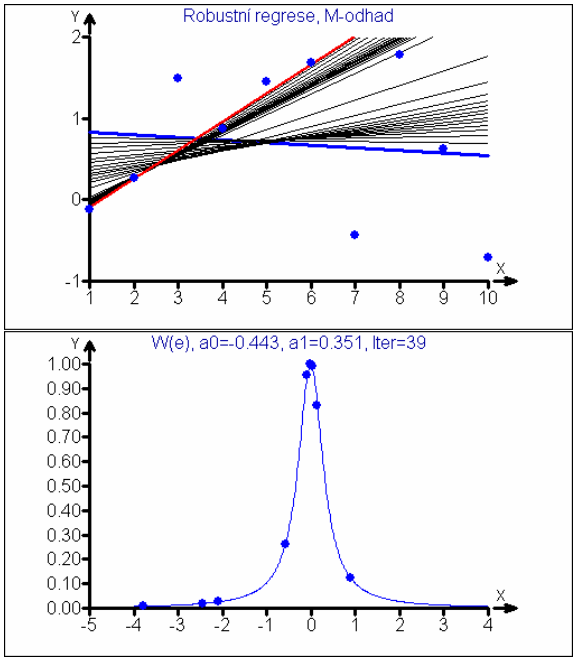
```
/******** Postupné iterační přímky a Weight Function: *******
N=10
xx=1:N
y=normalr(N)
n=count(y)
x=bind(ones(N),xx)
m=ncols(x)
plot(xx,y)
*/

w=ones(N) //Počáteční hodnoty vah (jednotkový vektor)
ap=rep(0,m)
xt=transp(x)
xtw=multdiag(transp(x),w) //Násobení diagonální maticí
xtwx=xtw#x
xtwx1=pinv(xtwx)
xtwy=xtw#y

a=xtwx1#xtwy // Odhady parametrů - obyčejné nejmenší čtverce (OLS), neboť
W=I
//a=normalr(2)
//a=vec(0,0.25)
aa=transp(a)
yp=x#a //Predikce
e=y-yp // Rezidua
s=mads(e) // Robustní směrodatná odchylka
w=weights(e,s) //Váhy z váhové funkce
w=w*w

plot(xx,y,main="Robustní regrese, M-odhad")
lineadd(a=a[1],b=a[2],width=3) // První přímka, metoda nejmenších čtverců
iter=0
while(and(gt(norm(a-ap),1e-6),lt(iter,200))) // Iterační cyklus
{
ap=a
xt=transp(x)
xtw=multdiag(transp(x),w) //Násobení diagonální maticí
xtwx=xtw#x
xtwx1=pinv(xtwx)
xtwy=xtw#y
a=xtwx1#xtwy // Odhad parametrů
yp=x#a
e=y-yp
s=mads(e)
w=weights(e,s)
w=w*w
lineadd(a=a[1],b=a[2],color=4) // i-tá přímka v iteračním procesu
iter=iter+1
}

lineadd(a=a[1],b=a[2],color=3,width=2) // Konečný M-odhad
plotadd(xx,y,color=0,ptsize=20)

// ********** // Graf váhové funkce se zobrazenými rezidui e:

xe=seq(-4,4,count=500)
we=weights(xe,s)
str="a0="+round(a[1],3)+", a1="+round(a[2],3)+", Iter="+iter
plot(xe,we,type=line,main="W(e), "+str)
plotadd(e,sqrt(w))
```



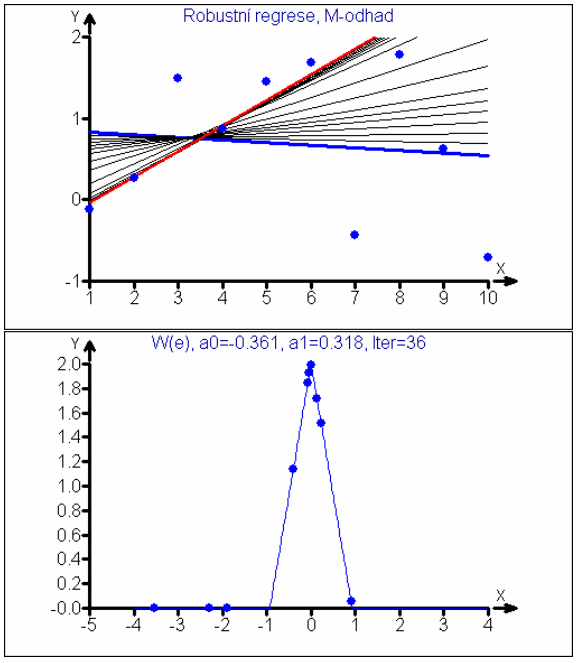
Obr. 35 Konvergence pro w(e) = exp(–|*e*|), Výsledné M-odhady: a_0=-0.404, a_1=0.337, počet iterací=38



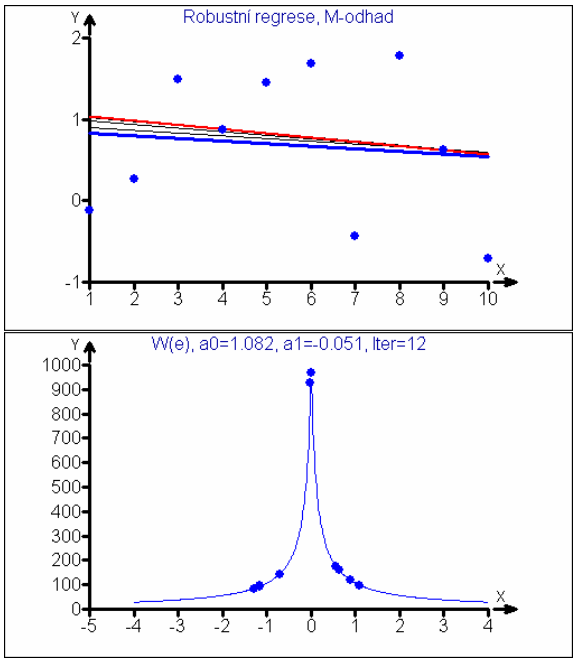
Obr. 36 Konvergence pro w(e) = exp(-(0.5*e2)), Výsledné M-odhady: a_0=-0.33, a_1=0.309, počet iterací=32



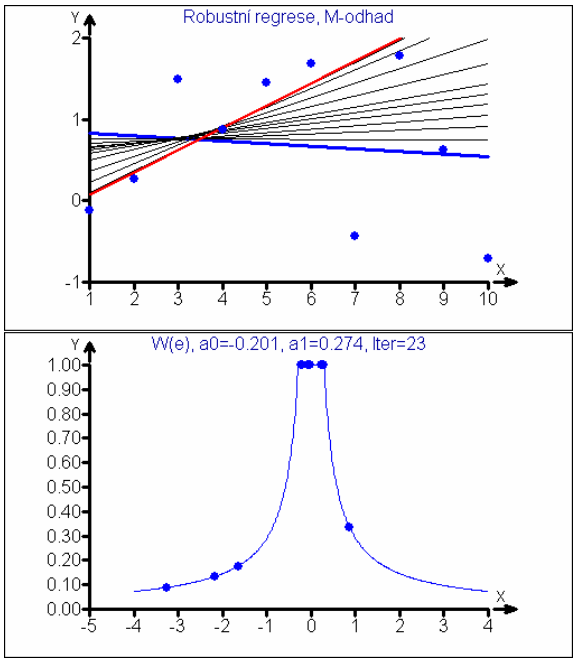
Obr. 37 Konvergence pro w(*e*) = 0.1/((0.5*e* +0.01)^2+0.1), Výsledné M-odhady: a_0=-0.443, a_1=0.351, počet iterací=39



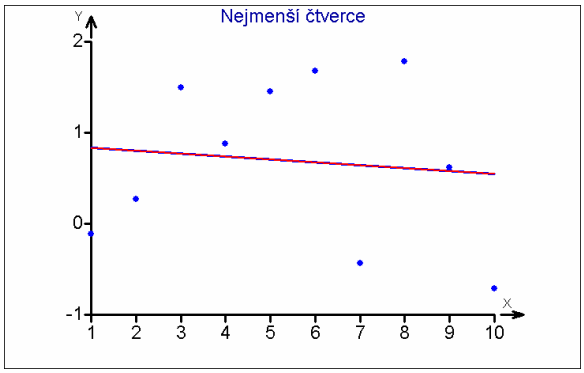
Obr. 38 Konvergence pro w(*e*) = 2 max(1–|*e*|;0) Výsledné M-odhady: a_0=-0.361, a_1=0.318, počet iterací=36



Obr. 39 Konvergence pro L_1 normu, w(e) = (1/(abs(0.01*e)+0.001)), Výsledné M-odhady: a_0=1.082, a_1=-0.051, počet iterací=12



Obr. 40 Konvergence pro Huberovu váhu, Výsledné M-odhady: a_0=-0.201, a_1=0.274, počet iterací=23



Obr. 41 Výsledné MNČ odhady: a_0=0.866, a_1=-0.032, w(*e*) = 1

5.2. Unikátnost M-odhadů v IRWLS regresi

Volbě konstant, které ovlivňují tvar funkce ρ, ψ, resp. w v M-odhadech (viz Obr. 9) se nevěnuje příliš pozornosti, obvykle se přebírají publikované empirické hodnoty bez jakékoliv kritiky, případně se simulačně volí konstanty, které dosahují stanoveného zlomku efektivity odhadu MNČ, například 0.95. Existují rovněž postupy volby ψ založené na funkci vlivu *IC* (*influence curve*) a extremalizaci Fisherovy informace $I(F(\boldsymbol{\theta}))$, kde lze ukázat, že efektivní M-odhad θ za předpokladu známé hustoty $f(x)$ se získá volbou

$$\psi(x) = -c\frac{f'(x)}{f(x)}, \qquad (38)$$

která vychází z požadavku na funkci vlivu ve tvaru

$$IC(x,F,T)=\frac{1}{I(F)}\frac{\partial \log f}{\partial \theta}, \qquad (39)$$

kde

$$I(F)=\int\left(\frac{\partial \log f}{\partial \theta}\right)^2 dF, \qquad (40)$$

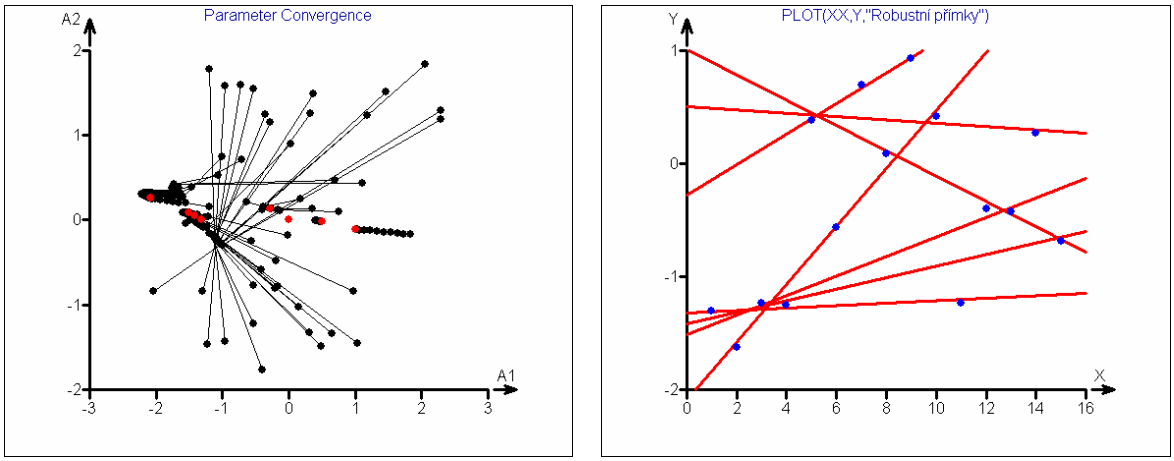
je Fisherova informace a funkce vlivu *IC* zavedená Hamplem se obyčejně vyjadřuje jako

$$IC(x,F,T)=\lim_{s\to 0}\frac{T\left[(1-s)F+s\delta(x)\right]-T(f)}{s}. \qquad (41)$$

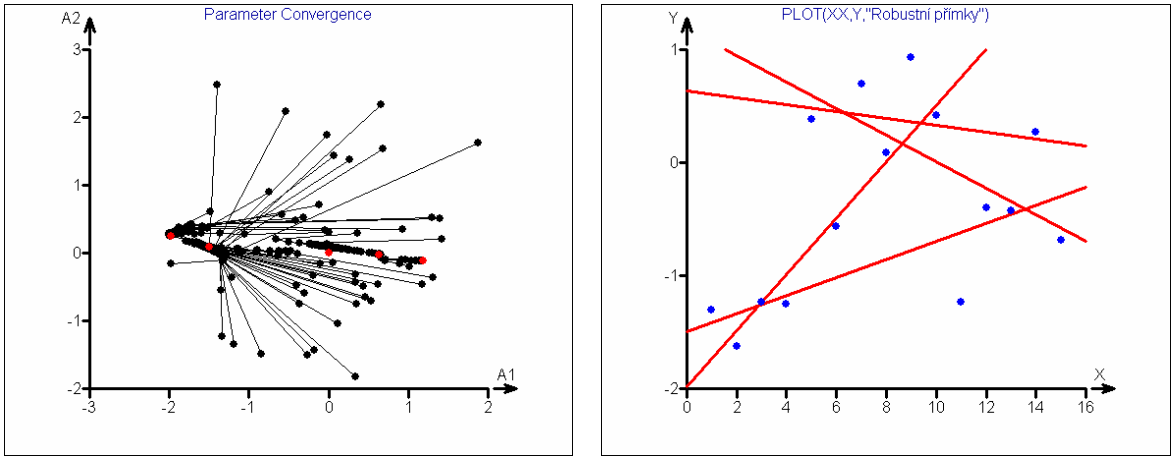
Pro normální hustotu pak dostáváme $\psi = x$ a $w = 1$ (metoda nejmenších čtverců, průměr), pro Laplaceovo rozdělení $\psi = x/|x|$ a $w = 1/|x|$ (minimum absolutních odchylek, medián), a podobně.

Vhodná volba „ladicích“ konstant *q* (obyčejně je jen jedna) má přitom zásadní vliv také na konvergenci a výsledek především v regresních úlohách. Pokud bychom v algoritmu (37) nevycházeli z odhadu MNČ, ale místo toho použili náhodné počáteční hodnoty $\mathbf{a}^0$ regresních parametrů z nějakého vhodně voleného dvourozměrného rozdělení, byly by takto získané M-odhady $\mathbf{a}^*$ závislé na $\mathbf{a}^0$ a také na konstantě *q* a tento algoritmus pak nemusí být jednoznačný. Pro příliš rychle klesající w(|*e*|) má $S_M(\boldsymbol{\theta})$ lokální minima, kterých může být až $(n^2 - n)/2$ a je obecně obtížné nalézt globální minimum, zvláště při větších hodnotách *n*. Na Obr. 42 až Obr. 46 jsou znázorněné vlevo postupné konvergence parametrů během IRWLS z 50 náhodných počátečních odhadů [generovaných z rozdělení $\mathbf{a}^0 \sim N(\mathbf{0}, \mathbf{I})$] k výslednému M-odhadu (červené body) v parametrickém prostoru pro přímku $y = a_1 + a_2x$. Vpravo jsou odpovídající výsledné robustní přímky ve výběrovém prostoru. V jednotlivých grafech se přitom mění pouze konstanta $q \geq 0$ použité váhové funkce $w(e) = \exp(-qe^2)$ (Welsh). Pro velké hodnoty *q* končí proces IRWLS v různých lokálních minimech $S(\boldsymbol{\theta})$ v závislosti na počátečním odhadu a výsledný odhad je velmi robustní, s break-pointem i nad 50%. Při snižování q se snižuje počet lokálních minim $S(\boldsymbol{\theta})$ a současně klesá robustnost odhadu, až při $q \to 0$ dostáváme metodu nejmenších čtverců. Tuto zdánlivou nevýhodu lokálních minim s rychle klesající $w(|x|)$ je však možné využít pro generování více možných vektorů optimálních robustních parametrů $\boldsymbol{\theta}_k^*$, které mohou představovat fyzikálně korektní množinu řešení odpovídající *r* regresním modelům pro jediný experiment s *n* pozorováními, přičemž příslušnost *i*-tého měření ke *k*-tému modelu

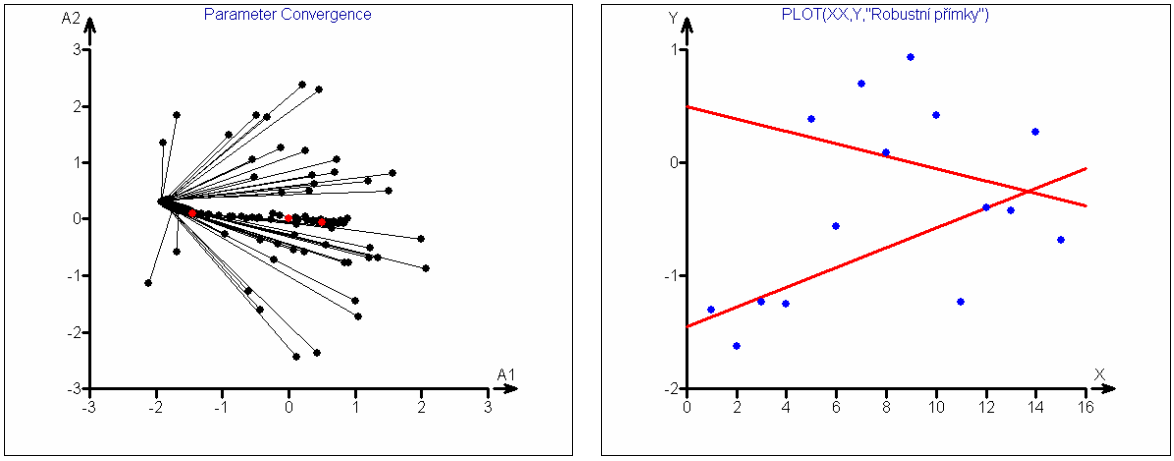
není předem známa. Případná interpretace takového výsledku je pak již věcí příslušného experimentátora.



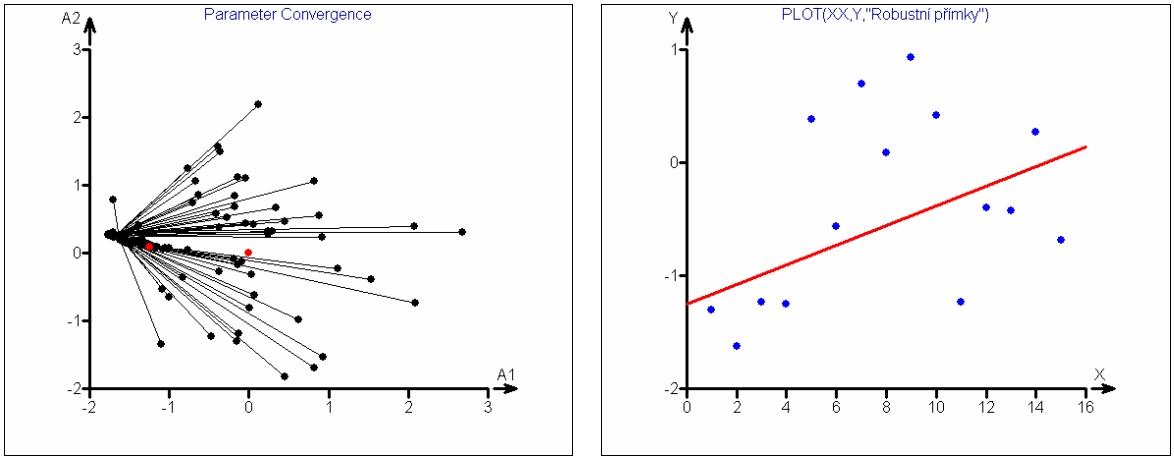
Obr. 42 $w(e) = \exp(-5e^2)$, 7 nalezených řešení



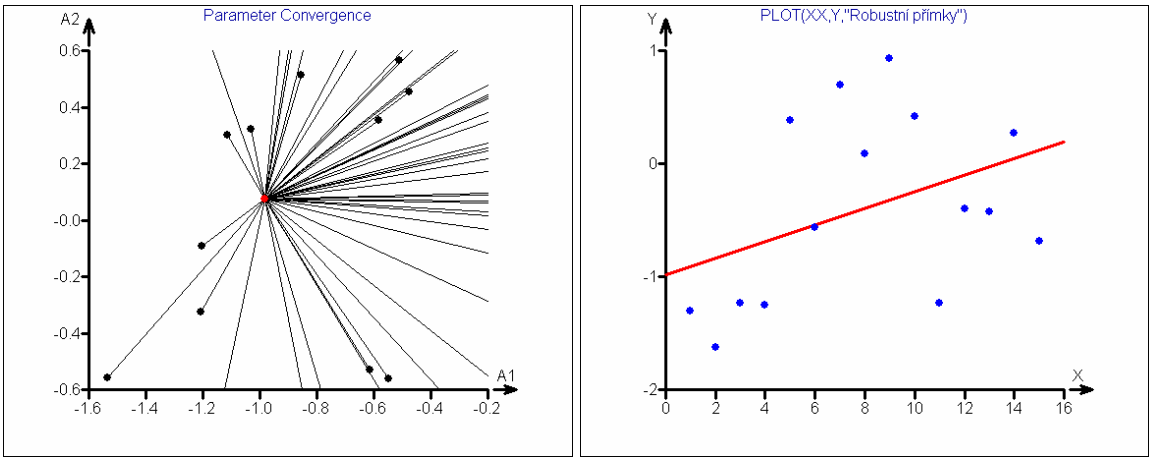
Obr. 43 $w(e) = \exp(-2e^2)$, 4 nalezená řešení



Obr. 44 $w(e) = \exp(-0.7e^2)$, 2 nalezená řešení

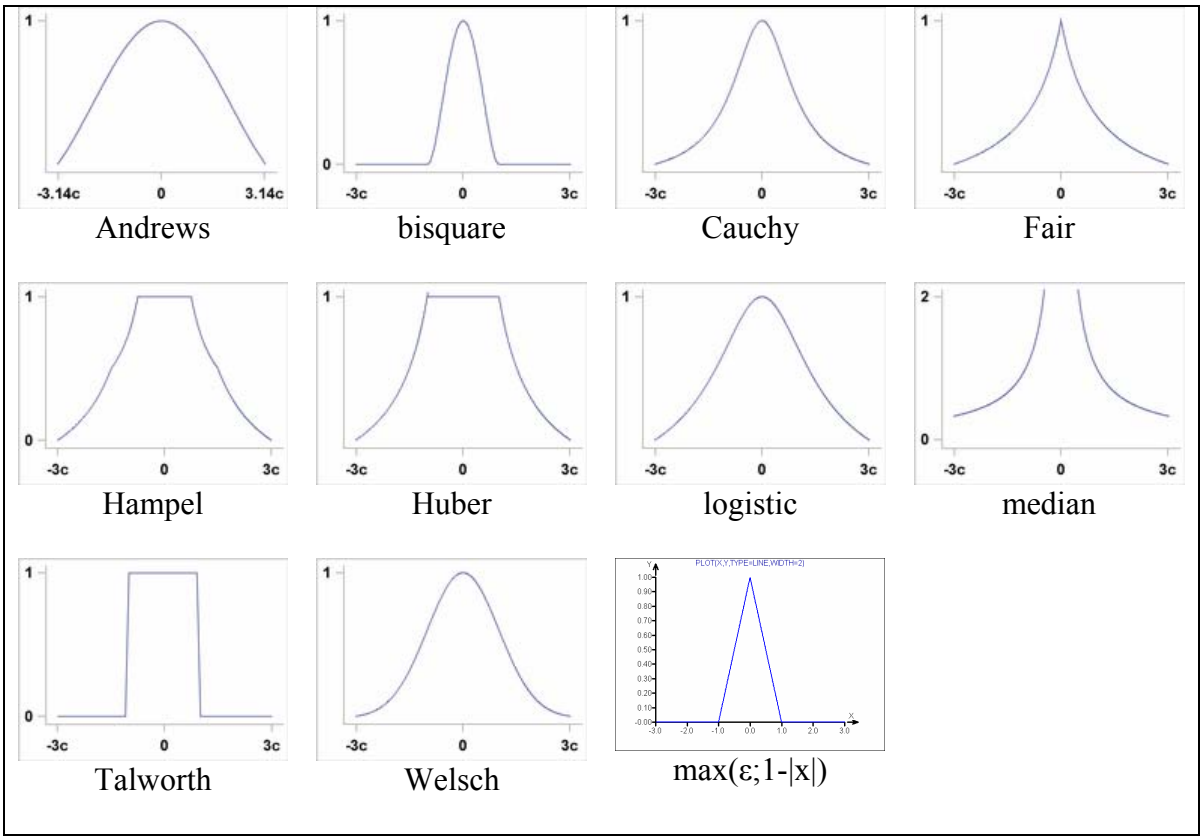


Obr. 45 $w(e) = \exp(-0.2e^2)$, 1 nalezené řešení



Obr. 46 $w(e) = \exp(-0e^2)$, Metoda nejmenších čtverců

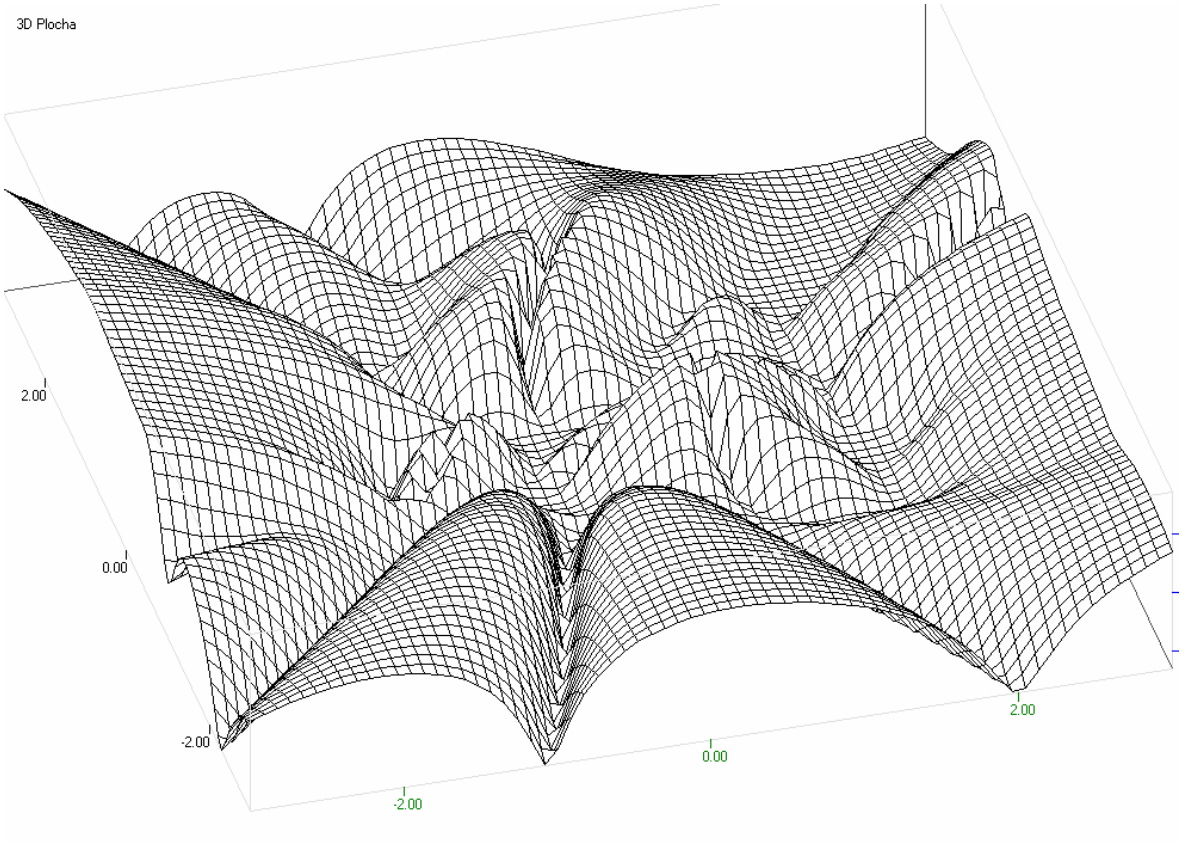
Na Obr. 48 až Obr. 52 jsou zobrazeny funkce $S(\boldsymbol{\theta})$ odpovídající příkladům Obr. 42 až Obr. 45, které ilustrují postupné zmírňování až zánik lokálních minim při snižování ladicí konstanty q ve váhové funkci a vysvětlují násobná řešení lineární regrese popsané výše. Příslušný skript použitý pro vytvoření 3d-ploch je uveden v tabulce



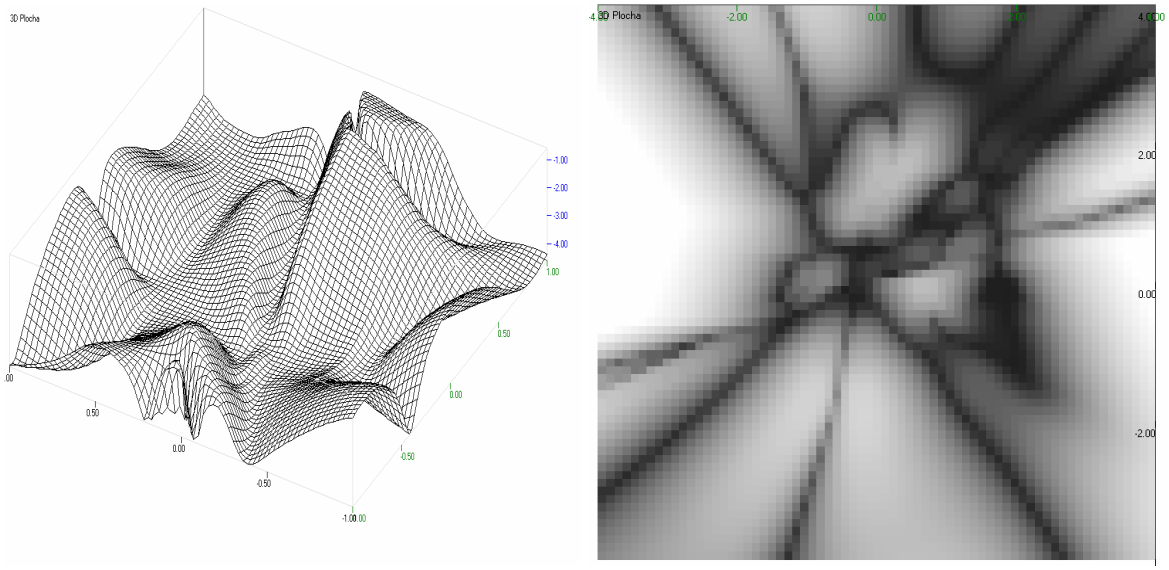
Obr. 47 Příklady váhových funkcí pro M-odhady ve statistickém systému SAS

Tab. 25 Skript v jazyce DARWin použitý pro generování grafů na Obr. 48 až Obr. 52

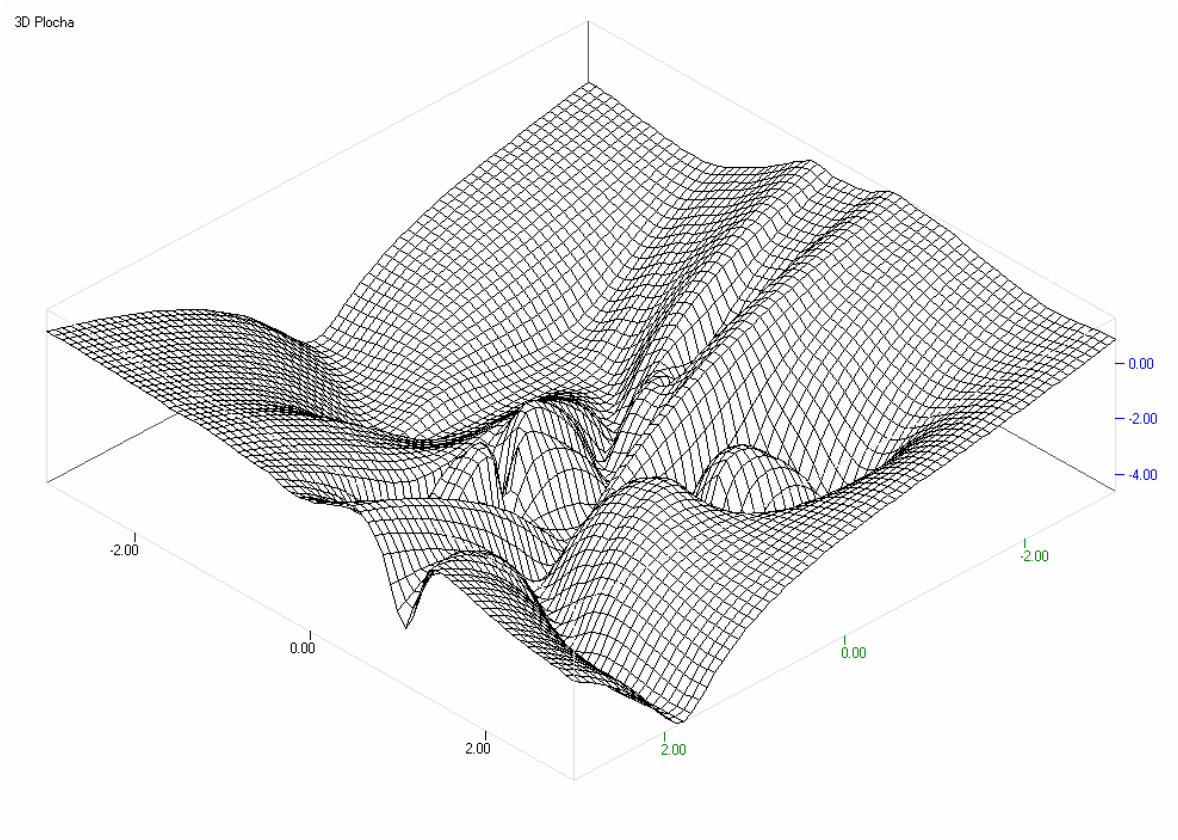
```
// M-Estimate Objective Function Plot

a1r=-2;a2r=2 // Meze pro a1
b1r=-0.4;b2r=0.4 // Meze pro a2
grid=50
delete(surf)
surf[grid,grid]=0 //Matice S(a1[i],a2[j])
a1=seq(a1r,a2r,count=grid) // Síť hodnot a1
a2=seq(b1r,b2r,count=grid) // Síť hodnot a2
for(i=1,grid)
{
for(j=1,grid)
{
a=vec(a1[i],a2[j])
yp=x#a // predikce
e=y-yp // rezidua
//s=sqrt(var(e))
s=mads(e) // Robustní směrodatná odchylka

w=weights(e,s) // Váhy w(e)

ssq=(sum(e*w)/sum(w))^2 //Vážený součet čtverců
surf[i,j]=ssq
}
}
surflog=ln(surf+0.01)
plot3dsurface(surf,xlim=vec(a1r,a2r),ylim=vec(b1r,b2r))
plot3dsurface(surflog,xlim=vec(a1r,a2r),ylim=vec(b1r,b2r))
```



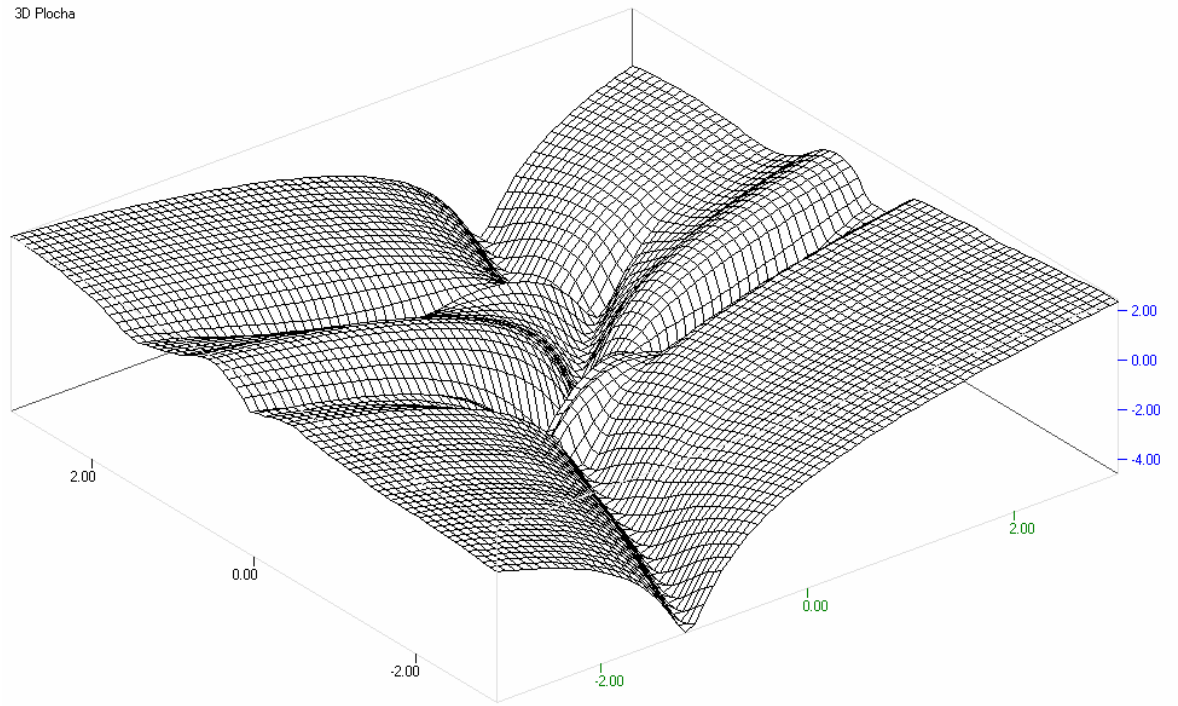
Obr. 48 ln(S+ε) při $w(e) = \exp(-5e^2)$



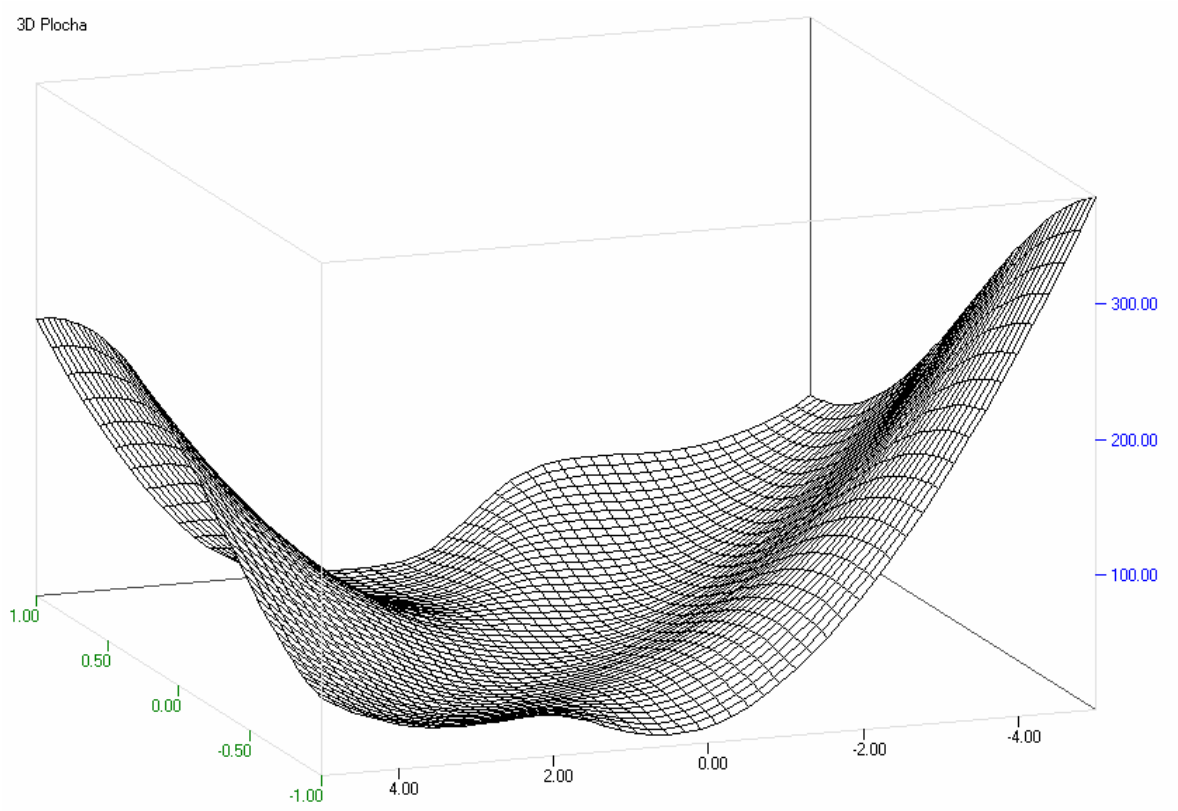
Obr. 49 ln(S+ε) při $w(e) = \exp(-5e^2)$, detail předchozího obrázku a odpovídající konturový graf, na němž jsou zřejmé tvary údolí a lokálních minim (tmavě).



Obr. 50 ln(S+ε) při $w(e) = \exp(-1e^2)$



Obr. 51 ln(S) při $w(e) = \exp(-0.2e^2)$



Obr. 52 *S* při $w(e) = \exp(-0.05e^2)$

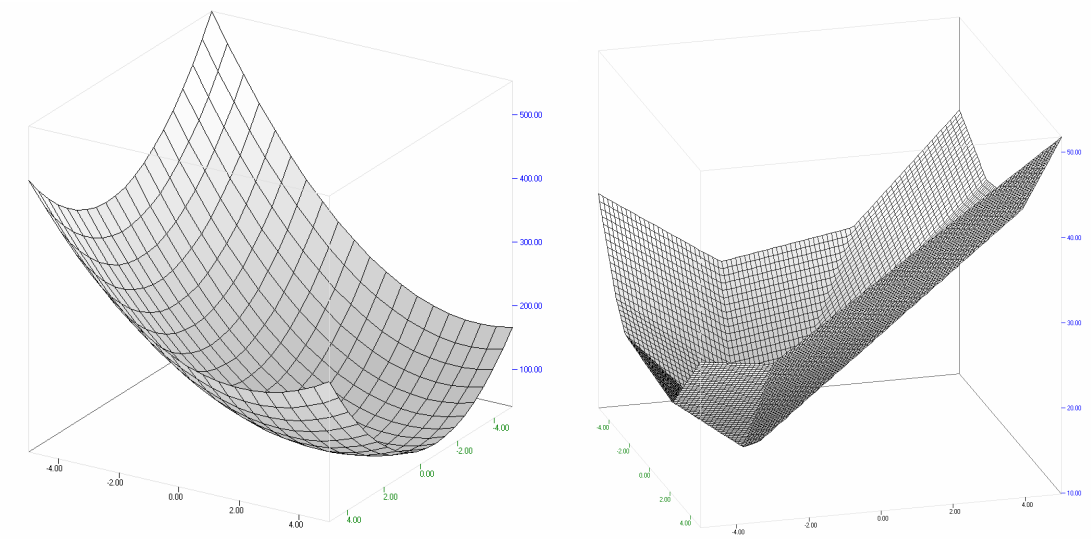
5.3. Lp regrese

Zobecněním kritéria minimálního součtu absolutních odchylek (5), strana 11 lze definovat robustní L_1-odhady parametrů regresního modelu jako m-rozměrný vektor $\boldsymbol{\theta}$ minimalizující pro $p = 1$

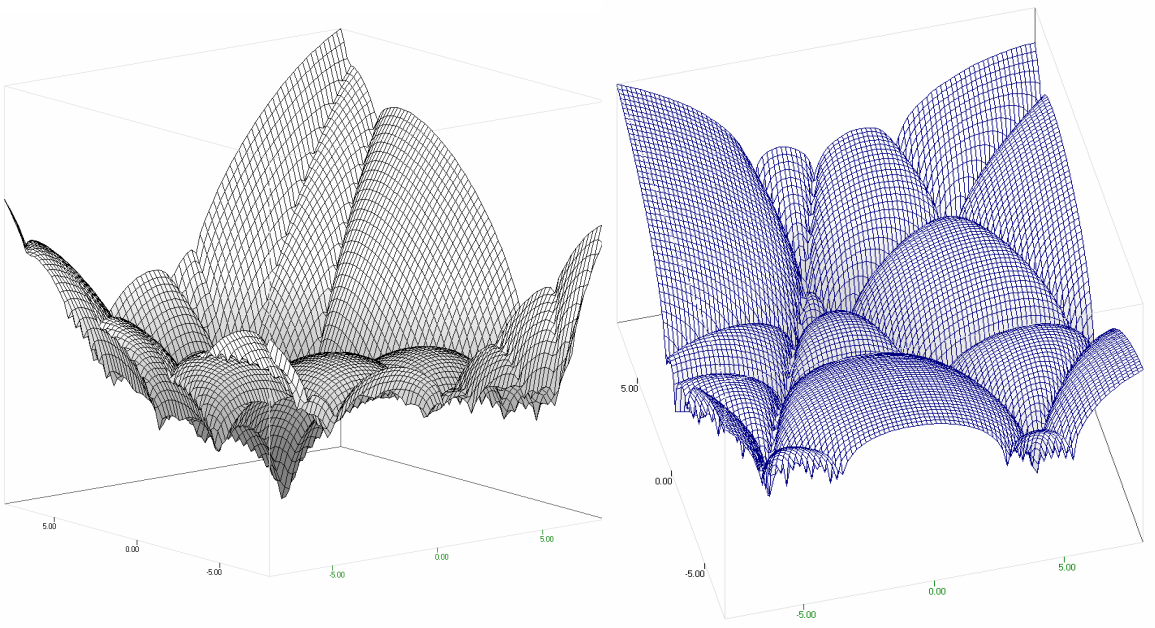
$$!\min_{\theta} S_{Lp}(\boldsymbol{\theta}) = \sum_{i=1}^{n} |e_i|^p = \sum_{i=1}^{n} |y_i - g(\boldsymbol{\theta}, \mathbf{x}_i)|^p . \qquad (42)$$

Pro lineární model $g(\mathbf{x}, \boldsymbol{\theta})$ má kritérium $S_{L1}(\boldsymbol{\theta})$ tvar m+1 rozměrného konvexního mnohostěnu s minimem ve vrcholu, nebo na 1- až m-rozměrném konvexním lineárním podprostoru v prostoru $(\boldsymbol{\theta}, S)$. V předchozím odstavci bylo ukázáno, že tento odhad je totožný také s M-odhadem při volbě $\rho = |x|$, případně $w = 1/x$ vycházejícím z Laplaceova rozdělení reziduí. V odstavci 3.1.1 byla diskutována rovněž možnost odhadů s $p < 1$. Následující Obr. 53 ilustruje tvar kritéria součtu čtverců (paraboloid) a součtu absolutních odchylek (polyedr) přímkového regresního modelu. Na Obr. 54 jsou ilustrovány tvary kritéria (42) pro $p = 0.5$ a dva parametry. Podle (38) by této podmínce odpovídala váhová funkce $w(x) = c\, x^{-3/2}$. Podobně jako u kritéria L_1 leží minimum v bodě funkce $S_{L0.5}(\boldsymbol{\theta})$, v nichž nejsou definovány derivace $\partial S_{Lp} / \partial \theta_i$. Tyto body leží v případě $m = 2$ obecně na n přímkách v parametrickém prostoru definovaných všemi dvojicemi parametrů $\boldsymbol{\theta}_k = (a_{1k},$

a_{2k}), $k = 1, .., (n^2 - n)/2$ přímek procházejících všemi dvojicemi bodů ($\mathbf{x}_i$, y_i), ($\mathbf{x}_j$, y_j), $i \neq j$. Hledání minima lze tak omezit na n jednorozměrných minimalizací podél těchto přímek. Na Obr. 55 a Tab. 26 je tato situace ilustrována na příkladu a na Obr. 56 je porovnání situace $L_{0.5}$ a MNČ regrese.



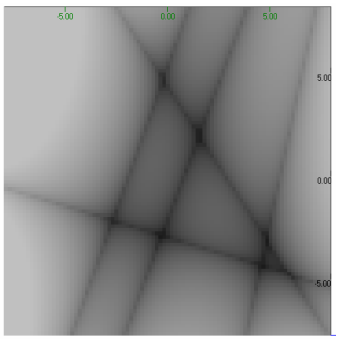
Obr. 53 Porovnání kriteriální plochy pro součet čtverců (L_2) a součet absolutních odchylek (L_1)



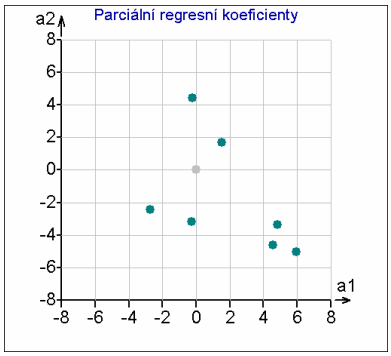
Obr. 54 Tvary kriteriální funkce pro kritérium $L_{0.5}$

Tab. 26 Příklad, data pro Obr. 55

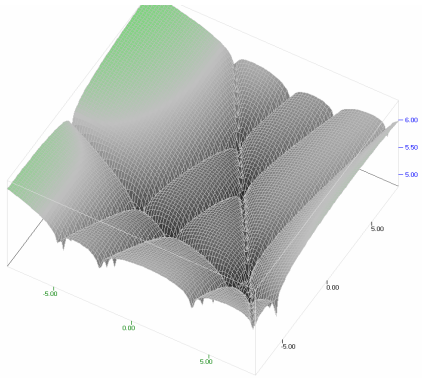
x_1	x_2	y
1.97	-0.74	1.77
-1.89	0.68	3.46
0.66	0.43	1.75
0.7	-0.15	3.89
-0.32	-1.06	3.44



A: Přímky v parametrickém prostoru, na nichž lze hledat minimum $S(\boldsymbol{\theta})$

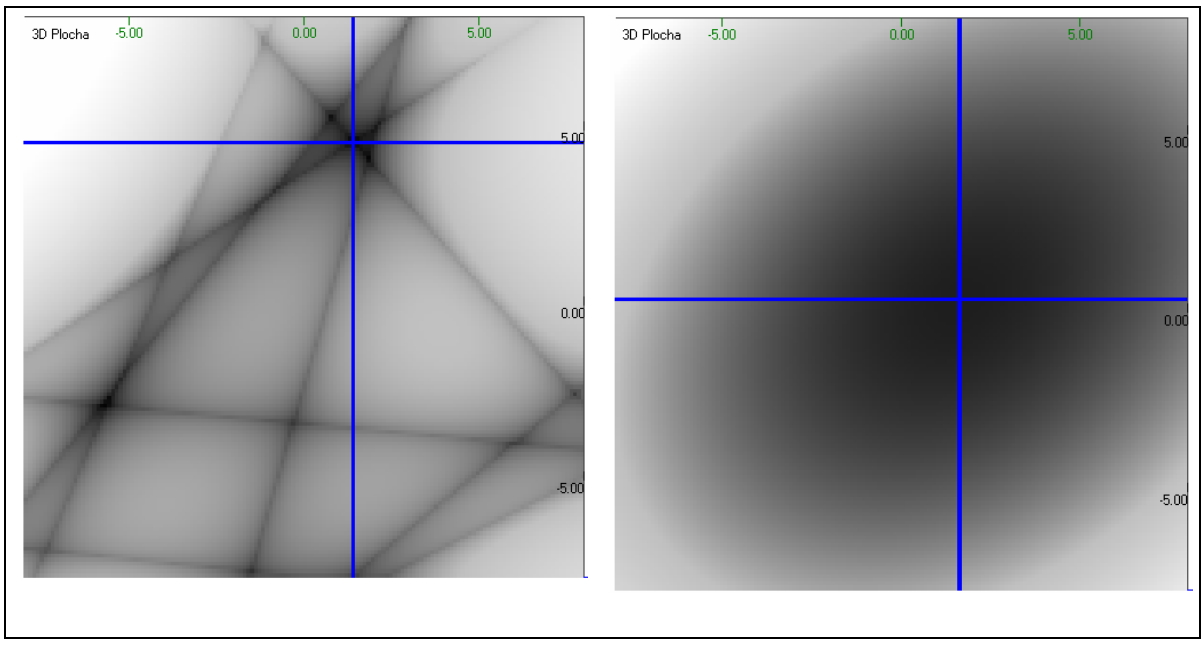


B: Parametry regresních přímek pro dvojice bodů



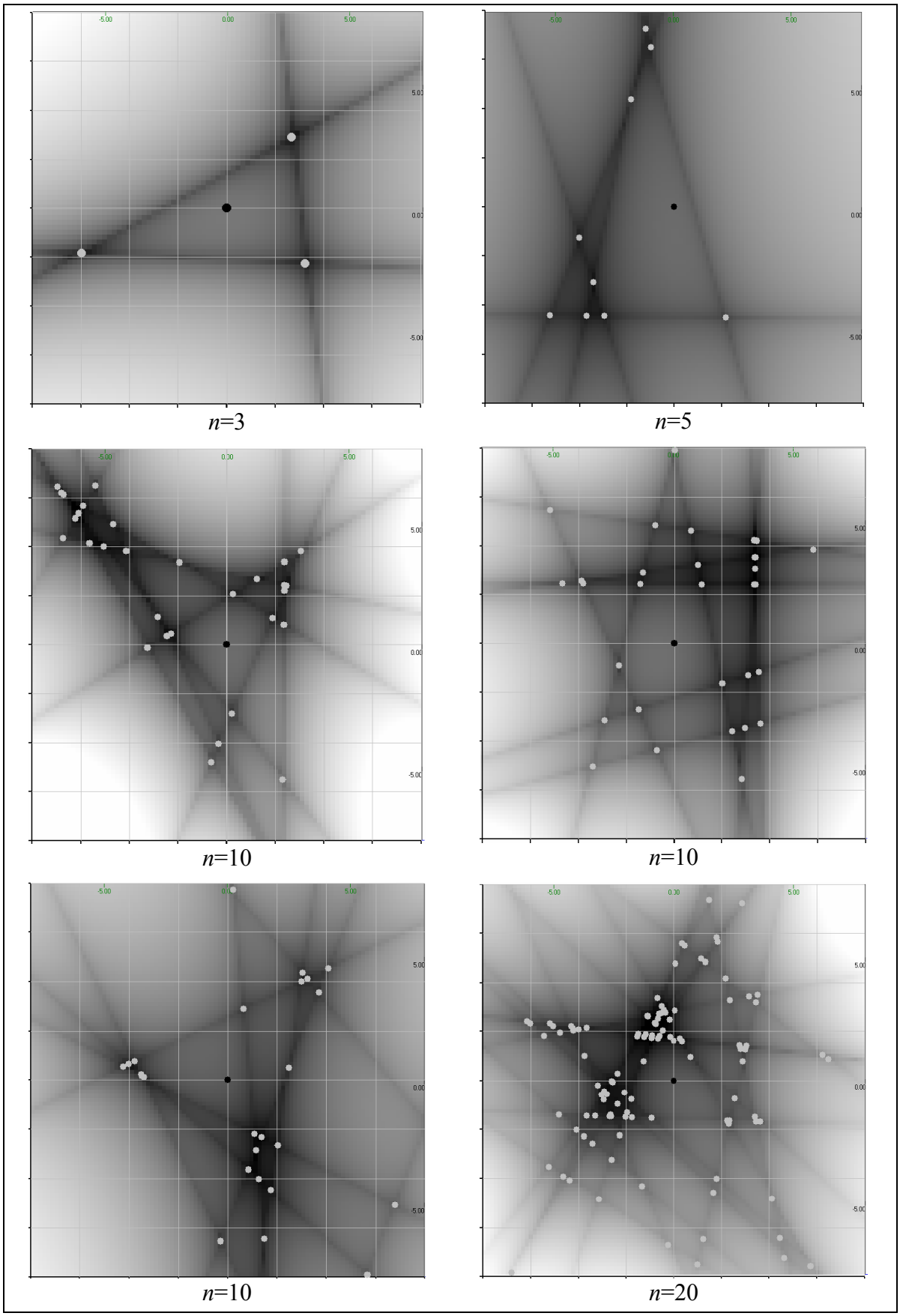
C: 3d znázornění $S(\boldsymbol{\theta})$

Obr. 55 Geometrické znázornění tvaru kritéria $L_{0.5}$



Obr. 56 Porovnání odhadu L0.5 a MNČ parametrů přímkového modelu pro stejná data

Odhad parametrů pro p=0.5 (QCExpert)	Odhad parametrů pro p=*2* (QCExpert)
Odhad parametrů a_1 = 1.45 a_2 = 4.87	Odhad parametrů a_1 = 1.64 a_2 = 0.73



Obr. 57 Příklady spojení Obr. 55 A a B v témže parametrickém prostoru, *m*=2 pro *n*=3, 5, 10, 10, 10, 20

Jak bylo zmíněno výše, body nespojitosti gradientu $S(\boldsymbol{\theta})$ leží v parametrické rovině a_1, a_2 na n přímkách s_1, s_2, .., s_n, jejichž úseky A_i a směrnice B_i jsou dány vztahy

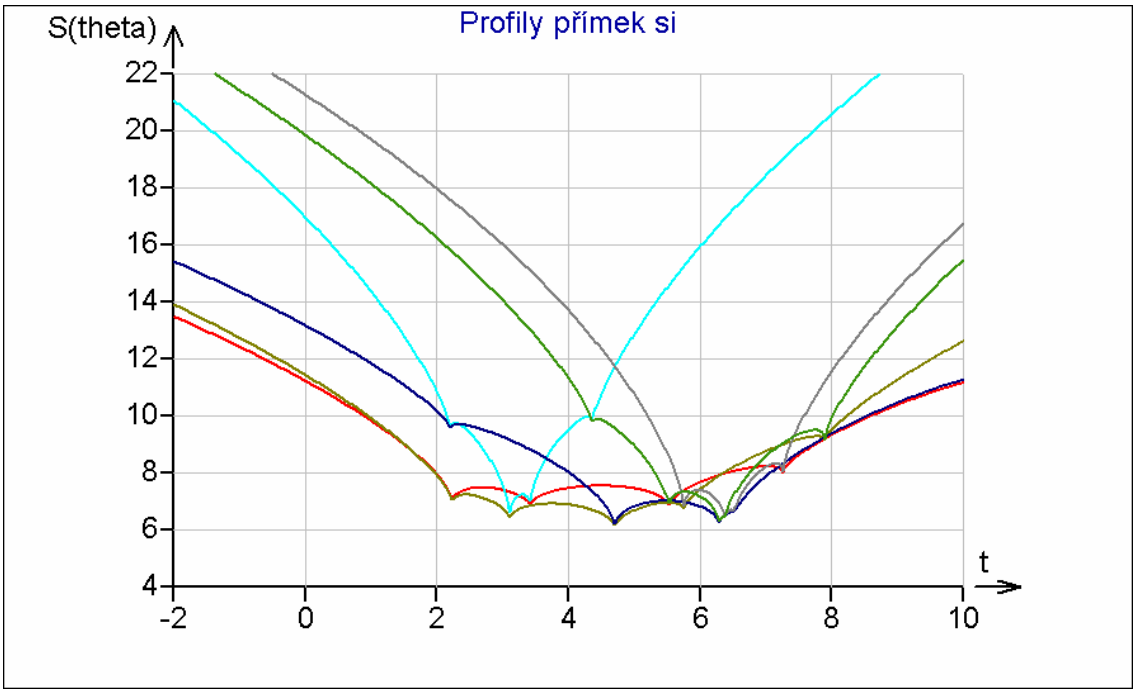
$$A_i = \frac{y_k - y_i}{x_k - x_i} - \frac{\left(y_i - \frac{x_i (y_k - y_i)}{x_k - x_i}\right)\left(\frac{y_j - y_i}{x_j - x_i} - \frac{y_k - y_i}{x_k - x_i}\right)}{-\frac{x_i (y_j - y_i)}{x_j - x_i} + \frac{x_i (y_k - y_i)}{x_k - x_i}} \qquad (43)$$

$$B_i = \frac{\frac{y_j - y_i}{x_j - x_i} - \frac{y_k - y_i}{x_k - x_i}}{-\frac{x_i (y_j - y_i)}{x_j - x_i} + \frac{x_i (y_k - y_i)}{x_k - x_i}} \qquad (44)$$

a jejichž rovnice po zjednodušení jsou

$$s_i : a_1 + a_2 x_i - y_i = 0\,. \qquad (45)$$

Tyto přímky jsou zobrazeny na Obr. 57 spolu s body odpovídajícími parametrům a_{1k}, a_{2k} přímek $y = a_{1k} + a_{2k}\, x$ procházejících všemi dvojicemi bodů (x_i, y_i), (x_j, y_j). Po parametrickém vyjádření s_i jako $(a_1, a_2) = f_i(t), t \in \mathbb{R}$ je možné snadno konstruovat průběhy lokálně minimálních hodnot $S(s_i)$ podél jednotlivých přímek s_i a rychle se orientovat ve tvaru minima, jak ilustruje Obr. 58 pro 6 bodů (x, y).



Obr. 58 Průřezové profily přímek s_i, graf S v závislosti na parametru t pro p=0.5

Tab. 27 Skript v jazyce DARWin použitý pro generování Obr. 53 až Obr. 57

```
//Konstrukce plochy účelové funkce Lp-regrese

n=6 // (počet bodů)
// >>>> Generování dat:
//x=bind(vec(1,1,1,1,1),vec(1,2,3,4,5)) // (Model y = a + bx)
x=bind(normalr(n),normalr(n)) // (Model y = ax1 + bx2, dobře podmíněný)
y=vec(0,2:n)+1
x=round(x,1)
y=round(y,1)

// >>> Konstrukce plochy účelové funkce Lp-regrese

poc=50 // (jemnost rastru)
ss=unit(poc) // (matice ss[poc x poc])
P=0.1 // (exponent p)
k1=-8
k2=8
l1=-8
l2=8
a1=seq(k1,k2,count=poc)
a2=seq(l1,l2,count=poc)
for(i=1,poc)
{
for(j=1,poc)
{
ss[i,j]=(sum(abs((x # vec(a1[i],a2[j]) - y))^P))
}
}
// >>> Zobrazení účelové funkce 3d:
plot3dsurface(ss, main="", xlim=vec(k1,k2), ylim=vec(l1,l2), angleX=0,
angleZ=270, colrange=vec(4,10,10), gridcolor=10)

// >>>> Regresní koeficienty pro všechny přímky procházející dvojicí
bodů:
plot(0,0,color=10) // černý bod (0,0)
for (i=1,n)
{
for (j=1,n)
{
if (gt(i,j))
{
aaM=inv(x[vec(i,j),])#y[vec(i,j)]
aa=aaM[1]
bb=aaM[2]
  plotadd(aa,bb,color=5)
}
}
}
```

5.4. Aplikace robustní regrese

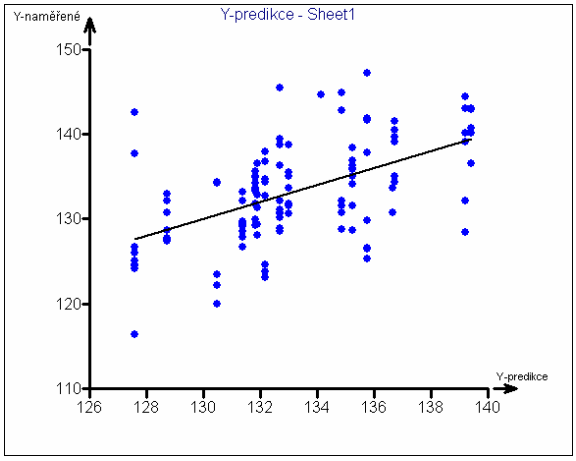
Na následujících dvou úlohách chceme porovnat použití, výsledky a interpretaci klasických nejmenších čtverců s robustním M-odhadem parametrů regresního modelu. Demonstrujeme zde možné výhody robustních metod v reálné situaci na reálných datech z mechanických zkušeben z roku 2005.

Úloha 1: Vliv legujících prvků a příměsí na pevnost (mez kluzu) hliníkové slitiny

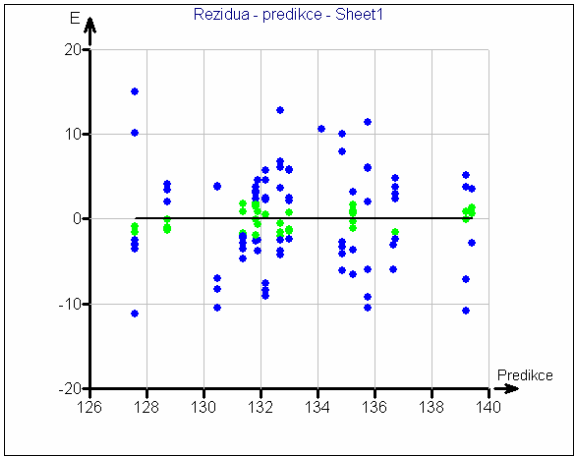
Tato úloha byla řešena pro Kovohutě Břidličná, a.s., data pochází z mechanických zkušeben. Mez kluzu R_e je mechanické napětí, při kterém elastická deformace materiálu přechází v plastickou, trvalou deformaci. Spolu s mezí pevnosti R_m je důležitým kvantitativním ukazatelem pevnosti. Cílem je vyšetřit lineární závislost R_e na kvantitativním zastoupení (koncentraci) kovových příměsí v hliníkové slitině. Lineární model je přijatelný vzhledem k malému rozsahu variability koncentrací. Regresní model lze symbolicky zapsat jako

$$R_e \sim Abs + Fe + Cu + Mn + Mg + Zn + Ti + Al + Cr + Ni$$

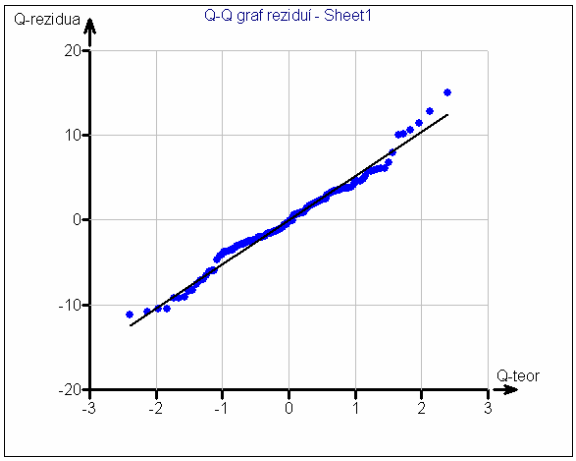
Použitím metody nejmenších čtverců (QCExpert) byly získány výsledky uvedené na Obr. 59 až Obr. 62 a v Tab. 28. Regresní koeficienty až na nikl jsou statisticky nevýznamné, což je v rozporu z očekáváním dle metalurgických a krystalografických modelů. V blízkém okolí modelu ($\pm 0.5s$) se nalézá třetina dat souboru.



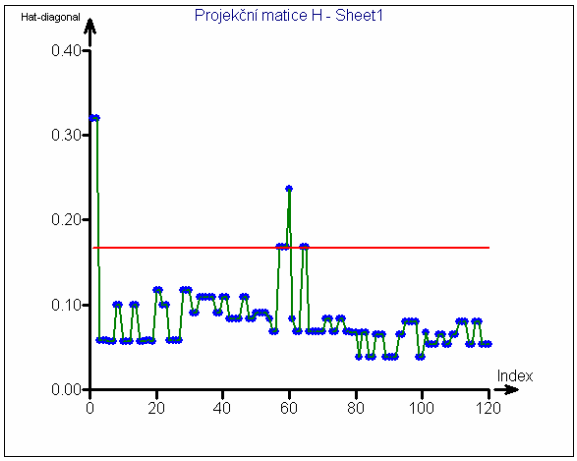
Obr. 59 Úloha 1, nejmenší čtverce, predikce



Obr. 60 Úloha 1, rezidua, 32% reziduí leží v intervalu ±2 (označeno zeleně)



Obr. 61 Úloha 1, normální QQ-graf



Obr. 62 Úloha 1, diagonála projekční matice

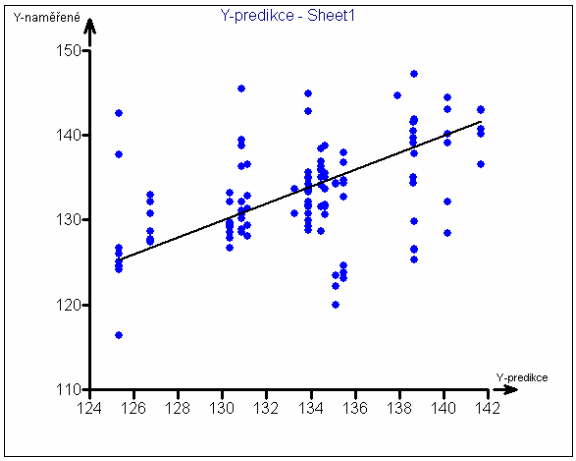
Tab. 28 Úloha 1, metoda nejmenších čtverců, výsledky regrese

Analýza rozptylu			
Průměr Y :	133.08667		
Zdroj	Součet čtverců	Průměrný čtverec	Rozptyl
Celková variabilita	4139.11867	34.49266	34.78251
Variabilita vysvětlená modelem	1185.5856	9.87988	9.9629
Reziduální variabilita	2953.53307	24.61278	24.81961
Hodnota kritéria F :	4.90615		
Kvantil F (1-alfa, m-1, n-m) :	1.96605		
Pravděpodobnost :	0.00002		
Závěr :	Model je významný		

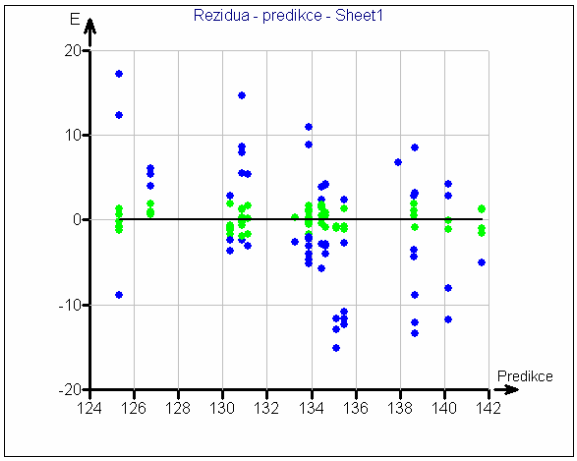
Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	p-hodnota	Spodní mez	Horní mez
Abs	-1763.13463	4798.08882	Nevýznamný	0.71398	-11271.82048	7745.55123
FE	88.13419	61.43608	Nevýznamný	0.15425	-33.6177	209.88609
CU	24.71153	190.79329	Nevýznamný	0.89718	-353.39599	402.81905
MN	60.14132	273.20119	Nevýznamný	0.82617	-481.27932	601.56195
MG	-1453.62205	1498.94066	Nevýznamný	0.33429	-4424.1706	1516.92651
ZN	929.63947	579.91818	Nevýznamný	0.11179	-219.62224	2078.90118
TI	-69.40153	57.67024	Nevýznamný	0.2314	-183.69041	44.88735
AL	18.58836	48.2293	Nevýznamný	0.70067	-76.99079	114.16751
CR	1610.48607	1391.59289	Nevýznamný	0.24966	-1147.3244	4368.29655
NI	-3167.65647	1319.77703	Významný	0.01807	-5783.14477	-552.16817

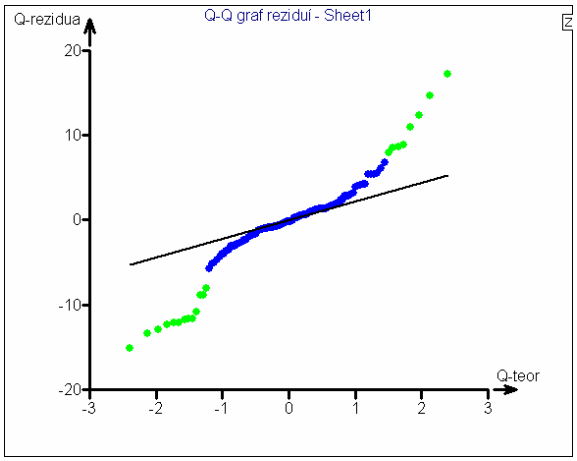
Použití robustního M-odhadu s váhovou funkcí typu exp($-e^2$) (Welsch, QCExpert) vedlo ke zcela odlišným výsledkům uvedeným v grafech na Obr. 63 až Obr. 66 a v Tab. 29. Pět z devíti regresních koeficientů je významných na hladině α=0.05 a jejich hodnoty potvrzují předpoklad vlivu železa, manganu, titanu a chromu na pevnost slitiny. V okolí modelu (± 0.5s) se nalézá polovina dat a navíc byly identifikovány vzorky, které tomuto modelu nevyhovují (na Obr. 65 jsou vyznačeny zeleně), které byly zřejmě chybně označeny, nebo změřeny. Použití robustního M-odhadu má tedy dva pozitivní praktické výsledky proti klasickým nejmenším čtvercům: nalezení smysluplného modelu a identifikace chybných (odlehlých) dat.



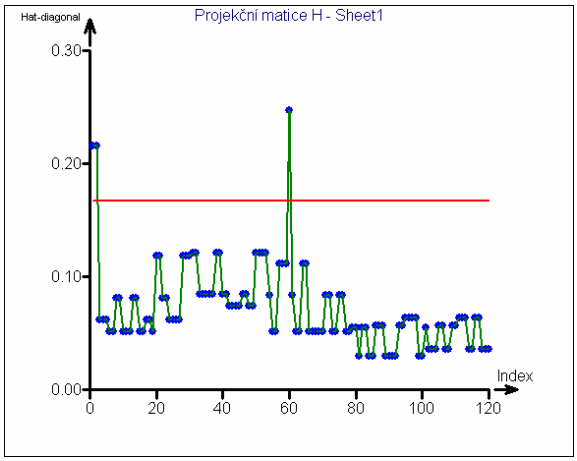
Obr. 63 Úloha 1, robustní M-odhad, predikce



Obr. 64 Úloha 1, robustní M-odhad, graf reziduí, 51% reziduí leží v intervalu ±2



Obr. 65 Úloha 1, robustní M-odhad, normální QQ-graf s vyznačenými vybocujícími hodnotami



Obr. 66 Úloha 1, robustní M-odhad, diagonála projekční matice

Tab. 29 Úloha 1, metoda M-odhad, výsledky regrese

Analýza rozptylu			
Průměr Y :	133.08667		
Zdroj	Součet čtverců	Průměrný čtverec	Rozptyl
Celková variabilita	4139.11867	34.49266	34.78251
Variabilita vysvětlená modelem	3612.27041	19.17966	19.18262
Reziduální variabilita	526.84826	29.41105	29.49999
Hodnota kritéria F :	83.80017		
Kvantil F (1-alfa, m-1, n-m) :	1.96605		
Pravděpodobnost :	4.427349943E-045		
Závěr :	Model je významný		

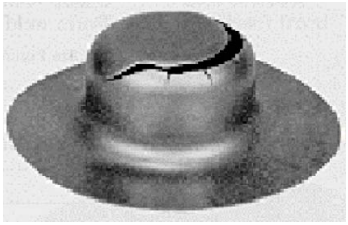
Odhady parametrů						
Proměnná	Odhad	Směr.Odch.	Závěr	p-hodnota	Spodní mez	Horní mez
Abs	-1819.43204	2163.30545	Nevýznamný	0.40215	-6106.59567	2467.73159
FE	86.391	27.36142	Významný	0.00205	32.16708	140.61491
CU	161.50729	82.29208	Nevýznamný	0.05222	-1.57629	324.59088
MN	633.68307	100.53241	Významný	6.203E-009	434.45144	832.9147
MG	-82.47096	629.75458	Nevýznamný	0.89605	-1330.49673	1165.55481
ZN	436.73048	234.41987	Nevýznamný	0.06513	-27.83468	901.29565
TI	-145.09481	24.68756	Významný	4.533E-008	-194.01976	-96.16986
AL	19.22456	21.74567	Nevýznamný	0.37859	-23.87025	62.31938
CR	-2365.50072	587.43446	Významný	0.0001	-3529.65795	-1201.3435
NI	-1182.96399	526.6218	Významný	0.02668	-2226.60479	-139.32318

Úloha 2: Erichsenova mechanická zkouška deformability (hlubokotažnosti)

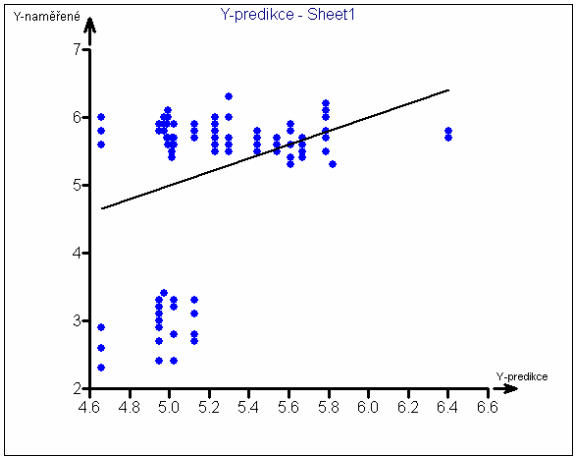
Erichsenova zkouška ([206] - [208]) se používá ke zjištění mechanických vlastností, především tažnosti, plochých materiálů. Měří se hloubka deformace plechu I_E v okamžiku protržení, viz Obr. 67. Předpokládal se lineární model

$$I_E \sim Abs + Mn + Mg + Ti,$$

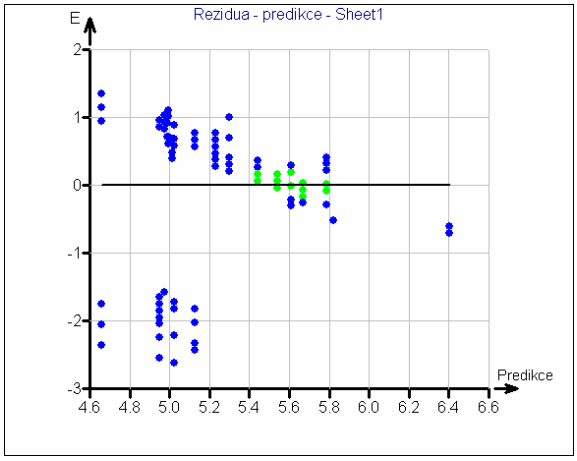
jehož parametry se opět odhadovaly klasickou a robustní lineární regresí. Výsledky klasické regrese jsou uvedeny na Obr. 68 až Obr. 71 a v Tab. 30, výsledky robustní regrese je na Obr. 72 až Obr. 75 a v Tab. 31. Cílem bylo ověřit, zda a jak ovlivňuje výsledek testu I_E obsah Mn, Mg a Ti. Výsledky jsou zřejmě výrazně ovlivněny skupinou 20 vybočujících dat patrnou v grafech. Zatímco metoda nejmenších čtverců vedla k bodovému odhadu parametrů Mn = -90, Mg = 370 a vliv Ti nevýznamný, robustní regrese dává zcela odlišné výsledky: Mn = 15, Mg = –80, Ti = 4, které odpovídají očekávanému pozitivnímu vlivu manganu a titanu a negativnímu vlivu hořčíku na I_E.



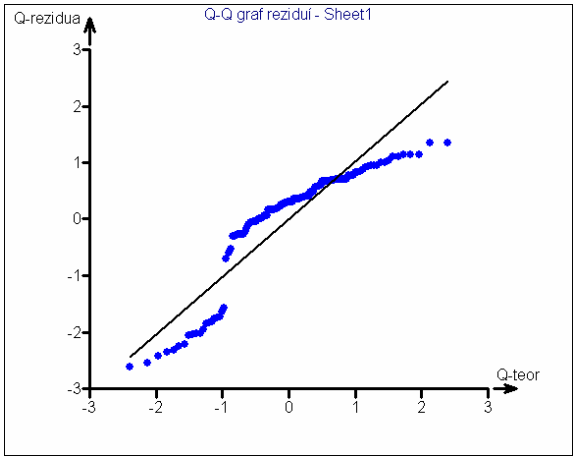
Obr. 67 Vzorek po Erichsenově zkoušce



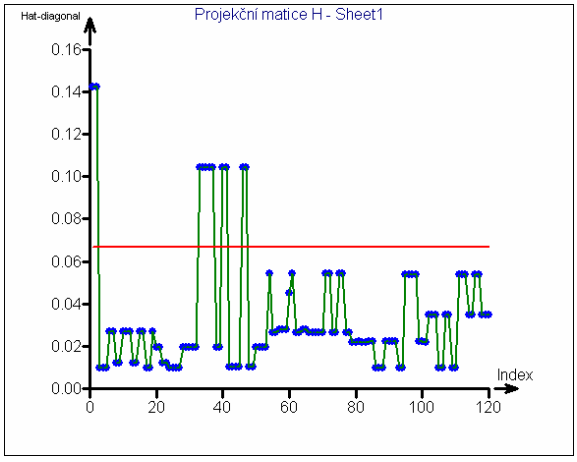
Obr. 68 Úloha 2, nejmenší čtverce, predikce



Obr. 69 Úloha 2, nejmenší čtverce, rezidua, 17% reziduí leží v intervalu ±0.2



Obr. 70 Úloha 2, nejmenší čtverce, QQ-graf pro normální rozdělení



Obr. 71 Úloha 2, nejmenší čtverce, diagonála projekční matice

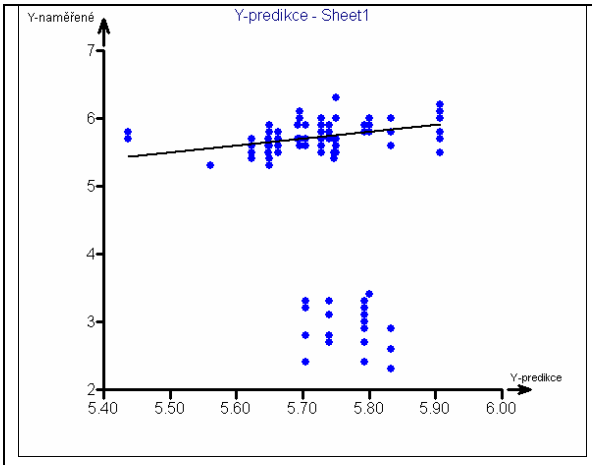
Tab. 30 Úloha 2, metoda nejmenších čtverců, výsledky regrese

Analýza rozptylu

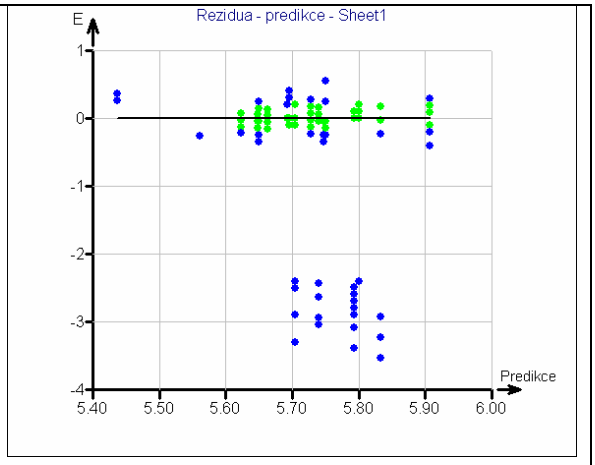
Průměr Y :	5.255		
Zdroj	Součet čtverců	Průměrný čtverec	Rozptyl
Celková variabilita	136.637	1.13864	1.14821
Variabilita vysvětlená modelem	15.33627	0.1278	0.12888
Reziduální variabilita	121.30073	1.01084	1.01933
Hodnota kritéria F :	4.8887		
Kvantil F (1-alfa, m-1, n-m) :	2.68281		
Pravděpodobnost :	0.00309		
Závěr :	Model je významný		

Odhady parametrů

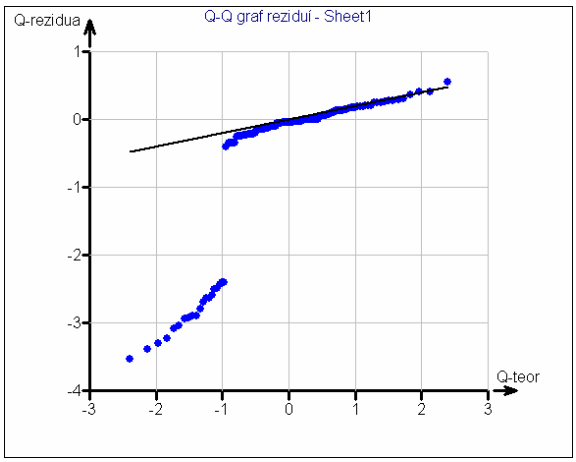
Proměnná	Odhad	Směr.Odch.	Závěr	p-hodnota	Spodní mez	Horní mez
Abs	4.56774	0.42615	Významný	0	3.72369	5.41179
MN	-90.15627	32.89749	Významný	0.00711	-155.31389	-24.99864
MG	369.8254	111.19316	Významný	0.00118	149.59335	590.05746
TI	11.37734	6.80573	Nevýznamný	0.09727	-2.10226	24.85694



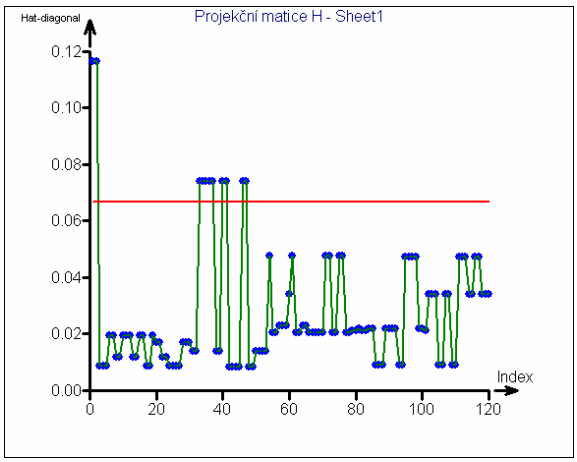
Obr. 72 Úloha 2, robustní odhad, predikce



Obr. 73 Úloha 2, robustní odhad, rezidua, 58% reziduí leží v intervalu ±0.2



Obr. 74 Úloha 2, robustní odhad, QQ-graf pro normální rozdělení



Obr. 75 Úloha 2, robustní odhad, diagonála projekční matice

Tab. 31 Úloha 2, metoda M-odhad, výsledky regrese

Analýza rozptylu

Průměr Y :	5.255		
Zdroj	Součet čtverců	Průměrný čtverec	Rozptyl
Celková variabilita	136.637	1.13864	1.14821
Variabilita vysvětlená modelem	132.04555	0.23059	0.00762
Reziduální variabilita	4.59145	1.39483	1.18165
Hodnota kritéria F :	1112.01552		
Kvantil F (1-alfa, m-1, n-m) :	2.68281		
Pravděpodobnost :	0		
Závěr :	Model je významný		

Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	5.71587	0.07718	Významný	0	5.56301	5.86872
MN	15.38364	5.77671	Významný	0.00884	3.94214	26.82513
MG	-81.40796	20.19975	Významný	0.0001	-121.41611	-41.39981
TI	3.66975	1.12223	Významný	0.00142	1.44703	5.89246

6. Modelování procesů pomocí dynamických modelů ANN-TS

6.1. Úvod

V této kapitole chceme naznačit na základě několika empirických simulačních studií možné využití neuronové sítě (ANN, Artificial Neural Network, [27] - [39]) pro modelování jednorozměrné časové řady (ANN-TS) a naznačit možnost konstrukce intervalů spolehlivosti těchto modelů a jejich předpovědi. Předpovídání pomocí neuronové sítě je intenzivně studovaná oblast. Jednou z hlavních nevýhod ANN-TS proti klasickým statistickým modelům, jako jsou modely typu ARIMA, SARIMA, GARCH, je obtížný, či nemožný odhad rozptylu predikce, konstrukce konfidenčních intervalů nebo testů. Dalším z problémů aplikace neuronové sítě je přeurčenost modelů, neboť počet parametrů neuronové sítě bývá řádově větší než u klasických modelů a díky jejich složité nelineární kovarianční struktuře mohou být modely ANN nestabilní. Dále naznačíme možnost využití především druhé jmenované nevýhody k částečnému odstranění nevýhody první. V textu budeme rozumět *i*-tým uzlem, neboli *i*-tým neuronem neuronové sítě funkci

$$z_i = \sigma\left(w_{i,0} + \mathbf{w}_i^{\,T}\mathbf{x}\right), \tag{46}$$

kde z_i je jednorozměrný výstup uzlu, $\mathbf{w}$ je vektor vah *i*-tého uzlu, x je vektor vstupních proměnných a $\sigma(x)$ je sigmoidální aktivační funkce

$$\sigma(x) = \frac{1}{1+e^{-x}}. \tag{47}$$

Hodnoty váhových koeficientů $\mathbf{w}_i$ u všech neuronů lze považovat za parametry modelu ANN-TS. Typickým důsledkem zmíněné přeurčenosti modelu ANN-TS je nestabilita parametrů, jejichž hodnoty jsou silně závislé na počátečním odhadu, kterým je náhodný vektor s rovnoměrným rozdělením v intervalu (-1, 1). Opakovaná optimalizace $\mathbf{w}_i$ tak dává pro stejná data vždy zcela rozdílné optimální vektory vah, i když samotný fit (proložení), vyjádřená například jako součet čtverců, je prakticky stejná.

Modely typu $\mathbf{y}_i \sim \Psi(\mathbf{x}_i)$, které dávají do relace hodnoty odpovídající stejnému časovému okamžiku t_i, se obyčejně označují jako statické. Index *i* může ovšem označovat například i stejné místo odběru vzorku, stejný živočišný druh, apod. V těchto modelech nehraje čas (místo, atd.) žádnou roli, není součástí modelu. Pokud zahrneme do modelu čas, označujeme modely jako dynamické (v případě zahrnutí souřadnic polohy mluvíme o prostorových modelech, *spatial models*, kde se bere v úvahu vzájemná poloha jednotlivých míst na mapě, nebo v prostoru). V případě dynamických modelů budeme pro jednoduchost

nejprve předpokládat, že měříme pouze jednorozměrné (skalární) hodnoty nějaké proměnné $y_i \in R^1$, které tvoří sloupcový vektor **y**. Jednotlivé prvky vektoru *n* naměřených hodnot **y** jsou data získaná v pravidelných časových intervalech (ekvitemporálně), takové posloupnosti dat se často říká časová řada. Dále budeme předpokládat, že naměřená hodnota y_i může souviset s hodnotami této proměnné naměřenými v předchozích okamžicích $y_{i-1}, y_{i-2}, \ldots, y_{i-m}$. Pak by bylo možné využít této informace k predikci budoucí neznámé hodnoty y_{i+1} z již naměřených hodnot $y_i, y_{i-1}, y_{i-2}\ldots$ K tomuto účelu by se data dala upravit vytvořením nových sloupců prediktorů posunutím původního sloupce vždy o jeden řádek, jak je naznačeno na obrázku níže. Závislost y_i na *m* předchozích měřeních by se dala vyjádřit například jednoduchým lineárním autoregresním modelem AR(*m*)

$$y_i = \alpha_1 y_{i-1} + \alpha_2 y_{i-2} + \ldots + \alpha_m y_{i-m} + \alpha_0 + \varepsilon = \sum_{j=1}^{m} \alpha_i y_{i-j} + \alpha_0 + \varepsilon ,$$

kde α_0 je absolutní člen regresního modelu související se střední hodnotou časové řady, zde $\alpha_0 = \left(1 - \sum_{i=1}^{m} \alpha_j\right) \bar{y}$. Pokud autoregresní model skutečně uspokojivě popisuje dynamiku časové řady, lze pomocí něj odhadnout (předpovědět) z *m*-tice posledních známých hodnot y_{i-j} budoucí hodnotu y_i. Protože u lineárních modelů lze spočítat i statistické vlastnosti odhadů parametrů a predikce, lze podobně jako u klasické lineární regrese konstruovat rovněž intervaly spolehlivosti předpovědi, což se využívá například u autoregresních regulačních diagramů, které mohou diagnostikovat odchylky od obvyklého pravidelného průběhu procesu, viz ilustrace níže.

Použije-li se místo autoregresního modelu neuronová síť se vstupy $x_{i-m}, x_{i-m+1}, .., x_{i-1}$ a výstupem x_i, je možné hledat model časové řady a ten pak případně využít pro předpověď budoucích hodnot $x_{i+1}, x_{i+2}, \ldots$. Teoretické články na téma neuronových časových řad jsou publikovány daleko méně často, než jiné modely ANN, viz [40] - [46], avšak aplikace ANN-TS jsou velmi rozšířené i bez hlubšího teoretického základu [47], [61]. Následující odstavce uvádějí simulační studie, které mohou přispět k náhledu na chování, vlastnosti a potenciální využití modelů ANN-TS.

Všechny výpočty a grafy byly realizovány autorem v prostředí jazyka DARWin a statistického systému QCExpert.

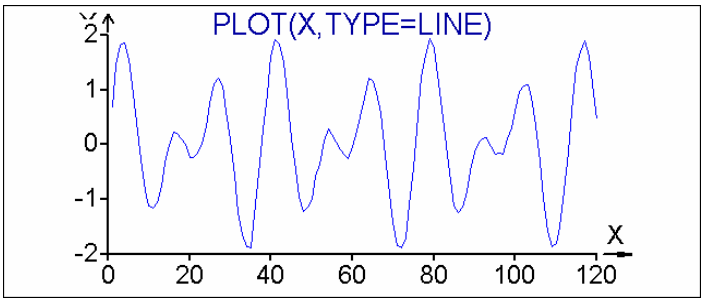
6.2. Simulační modelování periodického signálu

V následujících čtyřech simulacích ilustrujeme schopnost neuronové sítě (s architekturou (5,3), tedy 5 neuronů v první a 3 neurony ve druhé skryté vrstvě) modelovat funkční průběhy spojitých a nespojitých periodických signálů. Pomocí modelů 1 až 4 byla vygenerována data zatížená normálním iid šumem *N*(0, 0.04) a modelována

modelem ANN-TS(5,3). U jednotlivých modelů je uveden průběh teoretického signálu bez šumu, 120 vygenerovaných dat se šumem s fitem a 30-, resp. 20-krokovou předpovědí pomocí ANN-TS. Dále jsou uvedeny orientační grafy z analýzy hlavních komponent sestrojených z 500 opakování výpočtu optimálních vah se stejnými daty, ale různými (náhodnými) počátečními hodnotami vah (viz výše). Počet největších hlavních komponent naznačuje i reálnou dimenzionalitu modelu ANN a přibližně odpovídá počtu parametrů teoretického (v praxi obvykle neznámého) modelu. U každého modelu je uveden skript pro generování dat. Na konci odstavce je uveden skript v jazyce DARWin pro generování dat a optimalizaci ANN. V popisu je použita notace ANN $m(q_1, q_2)$, kde m je hloubka modelu, tedy počet hodnot časové řady použitých pro predikci následující hodnoty, q_1, q_2 jsou počty neuronů v první a druhé skryté vrstvě.

Simulační studie naznačuje, že i poměrně jednoduchý model ANN předpovídá poměrně spolehlivě hodnoty časové řady i v případě, kdy perioda je delší, než navzorkovaný úsek a také v případě, kdy jsou ve časové řadě nespojitosti a šum. Použité modely a algoritmy jsou uvedeny v tabulkách u jednotlivých příkladů.

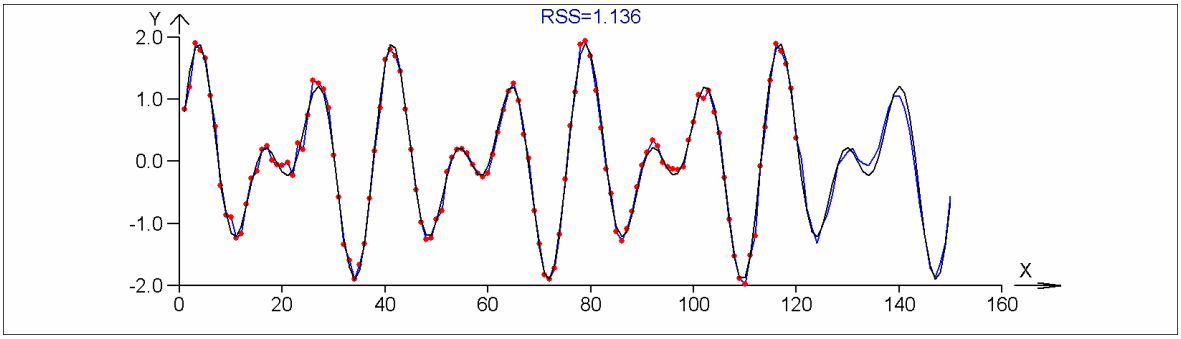
Model 1



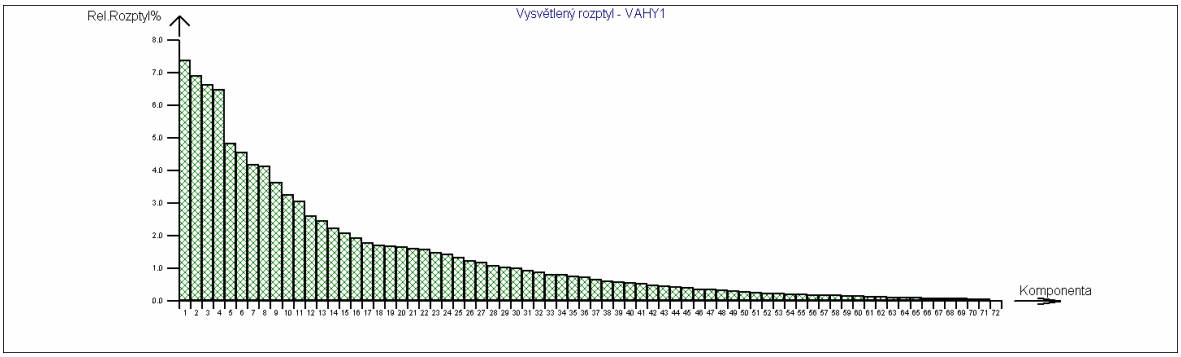
Obr. 76 Průběh teoretického modelu 1

Tab. 32 Použitý skript pro model 1

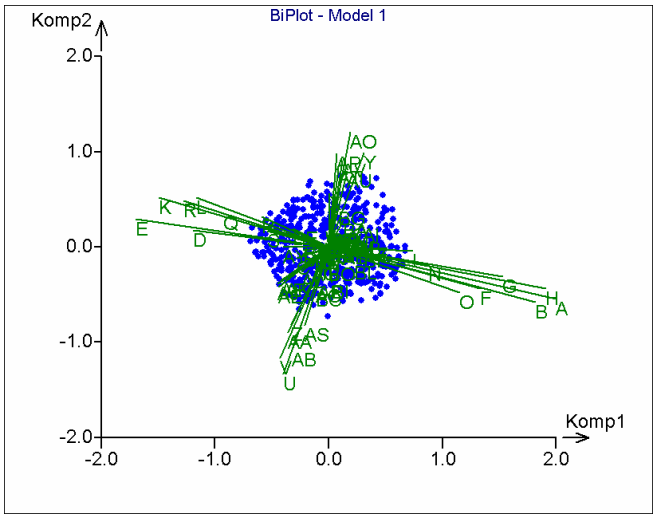
```
n=120 // Délka simulované řady
npred=30 // Počet předpovídaných hodnot
ndpth=9 // Hloubka modelu
narch=vec(5,3) // Architektura ANN
rnoise=0.2 // Směrodatná odchylka šumu
p1=2;p2=3
xi=1:n
x=sin(xi/p1)+sin(xi/p2)+normalr(n)*rnoise // Model 1
```



Obr. 77 Proložení a předpověď 30 hodnot pomocí ANN 8(5,3)

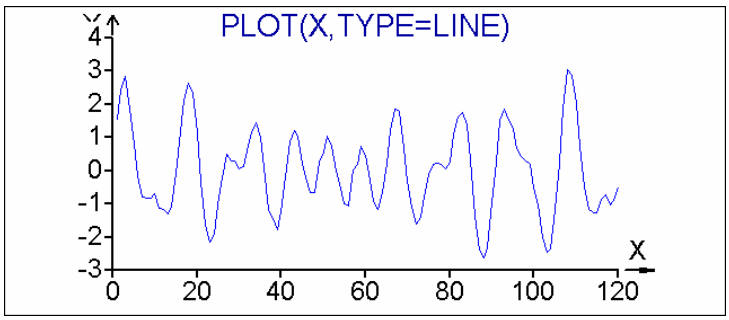


Obr. 78 Scree plot hlavních komponent naznačuje počet dimenzí 4



Obr. 79 Biplot pro parametry 500 opakovaně počítaných modelů naznačují počet dimenzí 4

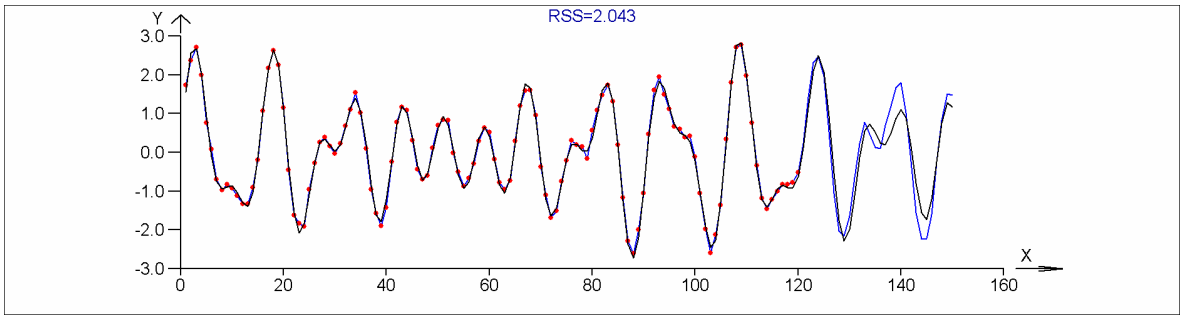
Model 2



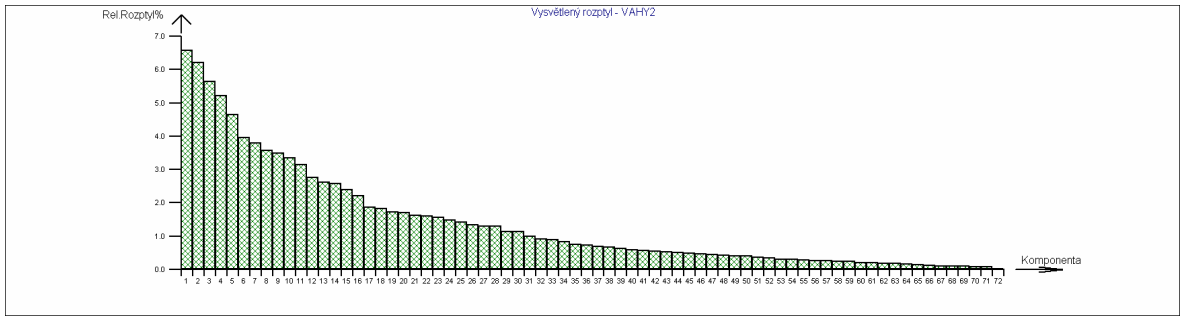
Obr. 80 Teoretickýprůběh modelu 2

Tab. 33 Použitý skript pro model 2

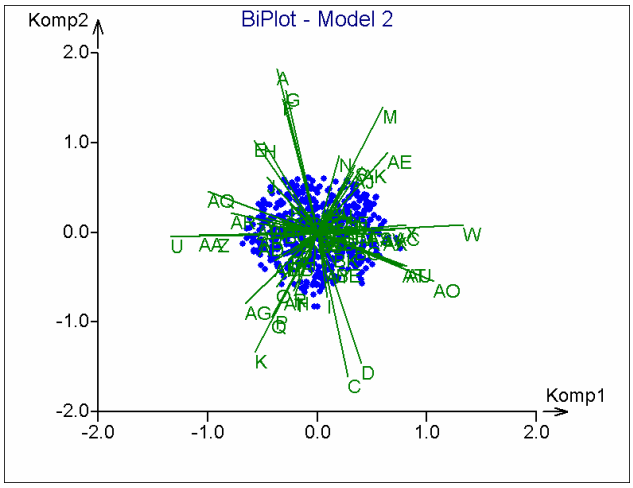
```
n=120 // Délka simulované řady
npred=30 // Počet předpovídaných hodnot
ndpth=9 // Hloubka modelu
narch=vec(5,3) // Architektura ANN
rnoise=0.2 // Směrodatná odchylka šumu
p1=1.3;p2=2.1;p3=2.4
x=sin(xi/p1)+sin(xi/p2) +sin(xi/p3)+normalr(n)*rnoise // Model 2
```



Obr. 81 Proložení a předpověď 30 hodnot pomocí ANN 8(5,3)

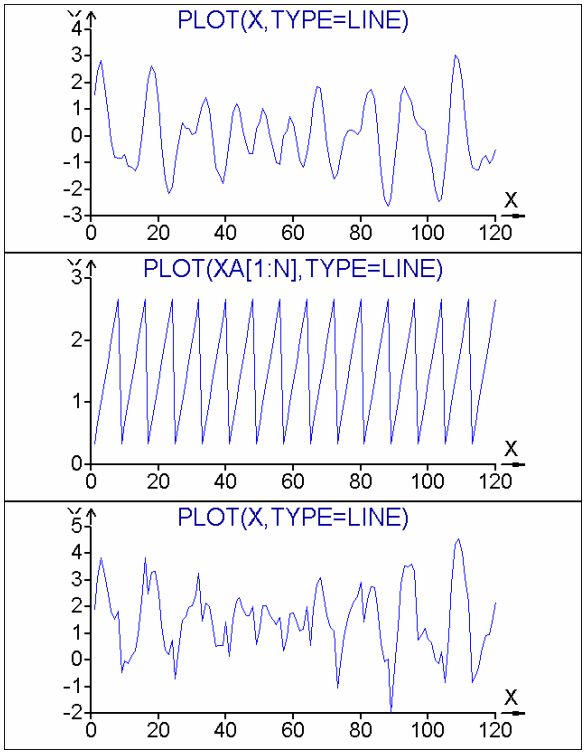


Obr. 82 Scree plot hlavních komponent, počet dimenzí 5, nebo 11



Obr. 83 Biplot pro parametry 500 opakovaně počítaných modelů č.2

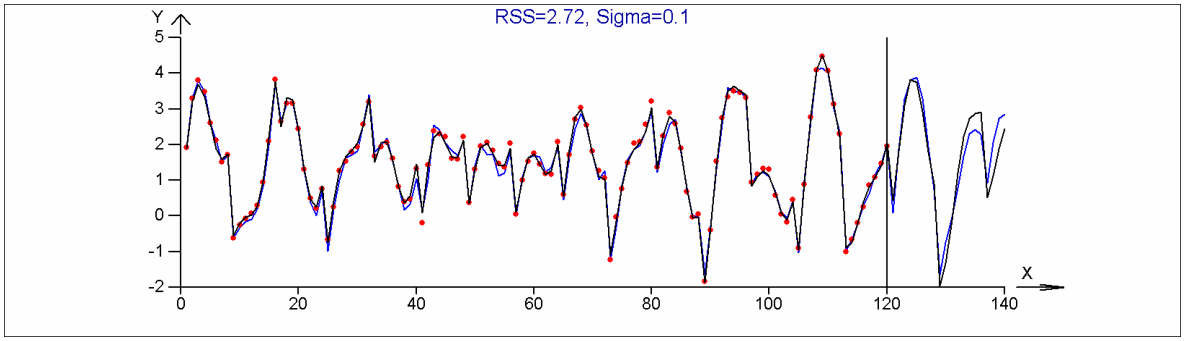
Model 3



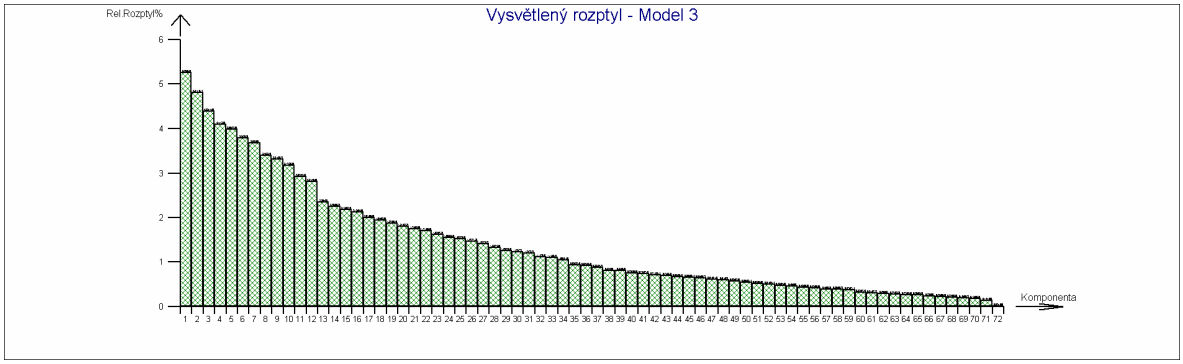
Obr. 84 Syntéza z periodického modelu a nespojité pily, výsledný teoretický průběh modelu 3

Tab. 34 Použitý skript pro model 2

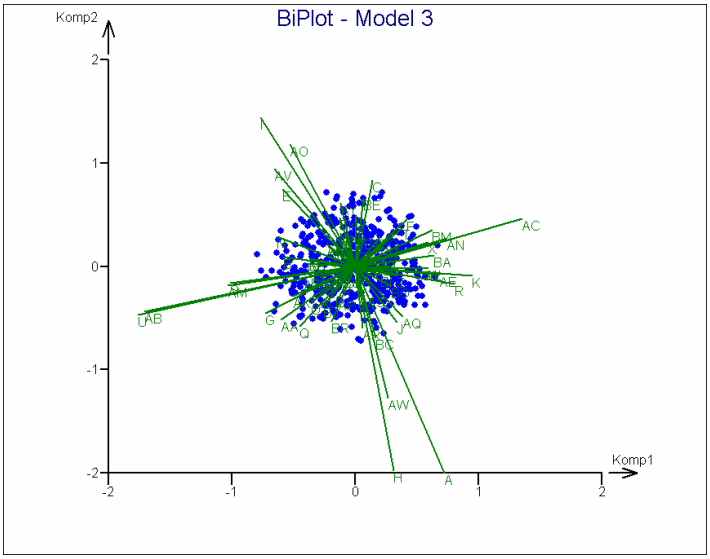
```
n=120                    // Délka simulované řady
npred=20                 // Počet předpovídaných hodnot
ndpth=9                  // Hloubka modelu
narch=vec(5,3)           // Architektura ANN
rnoise=0.2               // Směrodatná odchylka šumu
p1=1.3;p2=2.1;p3=2.4     // Hodnoty parametrů pro generaci funkčních hodnot
xi=1:n                   // Časová základna
x=sin(xi/p1)+sin(xi/p2)+sin(xi/p3)+normalr(n)*rnoise // Model 3
xa=rep(1:8,int(n/3))/3
x=x+xa[1:n]
```



Obr. 85 Proložení a předpověď 30 hodnot pomocí ANN 8(5,3)

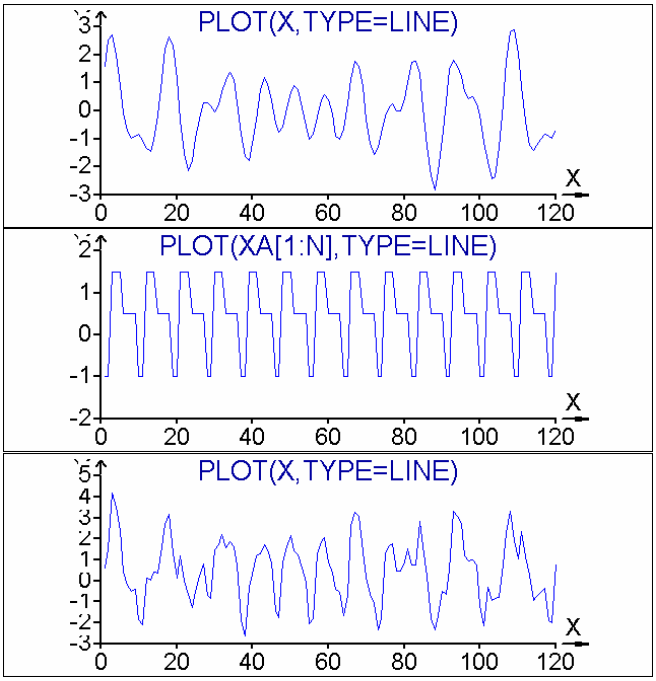


Obr. 86 Scree plot hlavních komponent, počet dimenzí 12

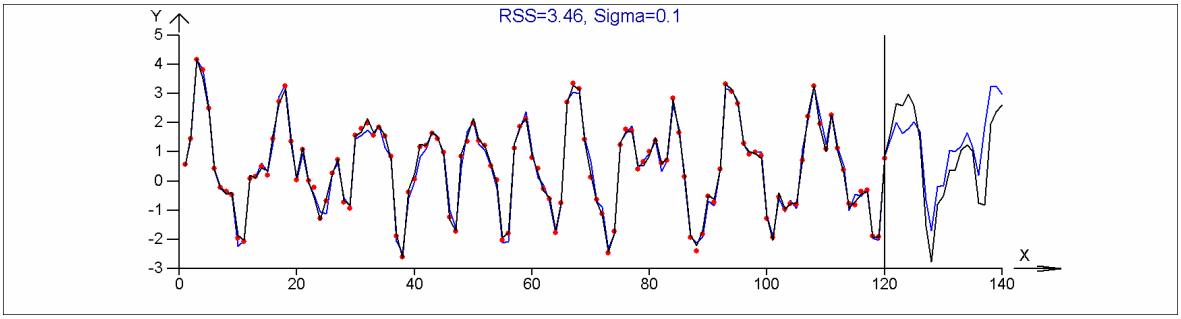


Obr. 87 Biplot pro parametry 500 opakovaně počítaných modelů č.3

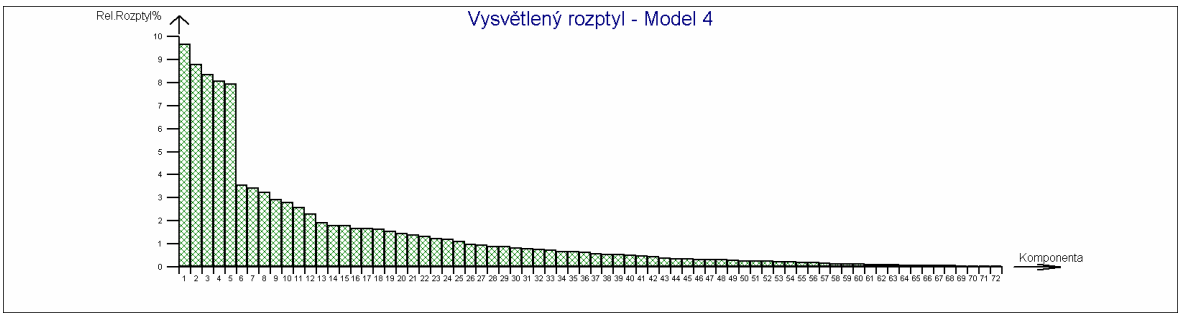
Model 4



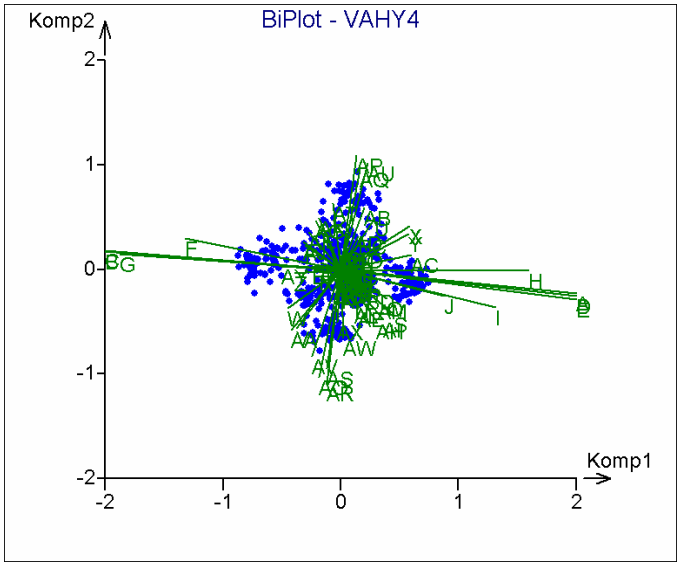
Obr. 88 Syntéza z periodického modelu stejného jako u modelu 3 a nespojité skokové funkce, výsledný teoretický průběh modelu 4



Obr. 89 Proložení a předpověď 30 hodnot pomocí ANN 8(5,3)



Obr. 90 Scree plot hlavních komponent, počet dimenzí 5



Obr. 91 Biplot pro parametry 500 opakovaně počítaných modelů č.3

Tab. 35 Použitý skript pro generování dat a optimalizaci ANN-TS modelu

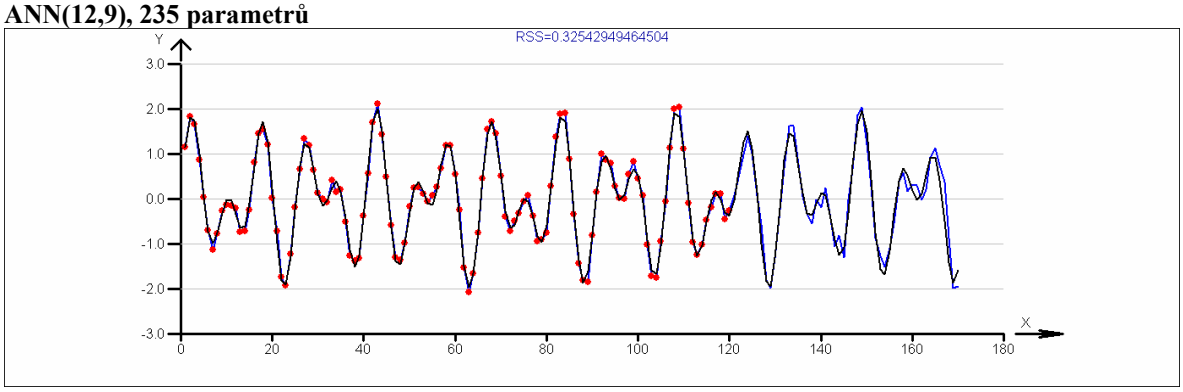
```
// Optimalizace ANN-TS:
n=120 // Délka simulované řady
npred=30 // Počet předpovídaných hodnot
narch=vec(5,3) // Architektura ANN
rnoise=0.1 // Směrodatná odchylka šumu
p1=1.3; p2=2.1; p3=2.4
ndpth=8 // Hloubka modelu
x=sin(xi/p1)+sin(xi/p2)+normalr(n)*rnoise // Model 1
//x=sin(xi/p1)+sin(xi/p2)+sin(xi/p3)+normalr(n)*rnoise // Model 2
//x=cusum(normalr(120)) /Random Walk
aa=nntimelearn(x,ModelDepth=ndpth,Layers=narch,Iterations=8000,Residuals=
1,grnet=1)
aap=nnpredict(x,model="aa",forecast=npred)
plot(vec(x,aap$forecast),type=line,main="RSS="+sum(AA$Residuals^2),width=
2)
plotadd(x,color=3,width=2)
x2i=1:(n+npred)
//x2=sin(x2i/p1)+sin(x2i/p2)+sin(x2i/p3) //Model 2
x2=sin(x2i/p1)+sin(x2i/p2) // Model 1
plotadd(x2,type=line,color=4,width=2)
```

6.3. *Vliv počtu parametrů na předpověd*

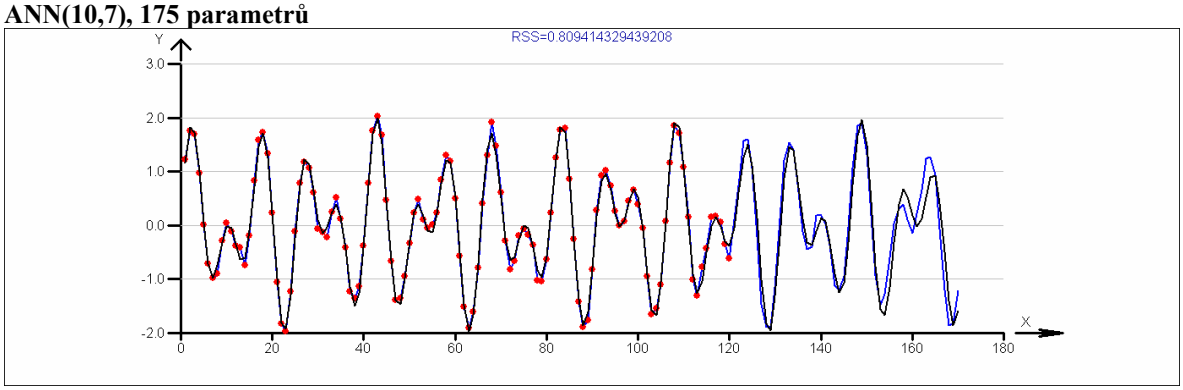
Tato simulační studie navazuje na předchozí odstavec. Modelují se v ní stejná data (až na šum) pomocí ANN-TS modelů s 2 skrytými vrstvami neuronů, které se liší složitostí, od modelu ANN 8,(12,9) s 235 parametry až po model ANN 8,(1,2) s 12 parametry. Studie ukazuje, že modely s minimálním počtem parametrů, ale i velmi přeurčené modely ANN-TS s počtem parametrů až dvakrát větším, než počet dat jsou stále použitelné pro modelování časové řady a mají spolehlivou předpověď. Simulace se provádějí na dvouparametrickém modelu 1 z předchozího odstavce. Použité modely a algoritmy jsou uvedeny u jednotlivých příkladů.

```
p1=1.3;p2=2.1
x=sin(xi/p1)+sin(xi/p2)+normalr(n)*rnoise
//Hloubka modelu q=8
```

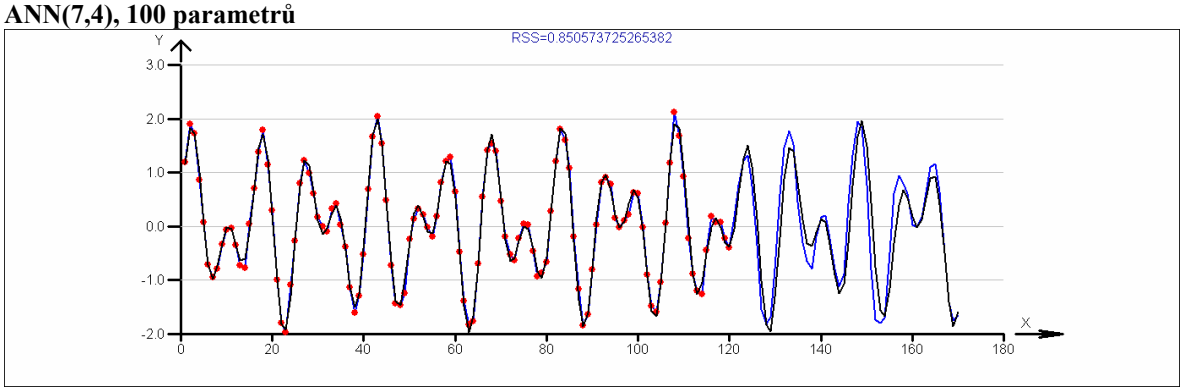
ANN(12,9), 235 parametrů



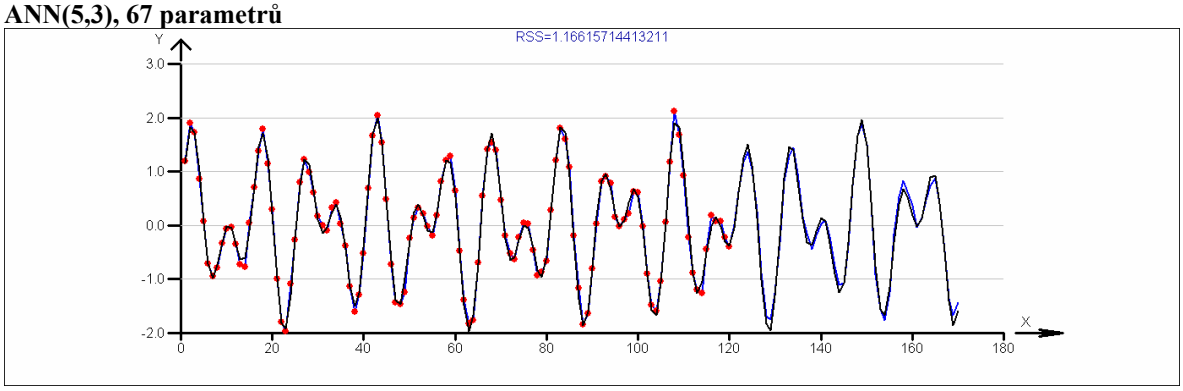
ANN(10,7), 175 parametrů



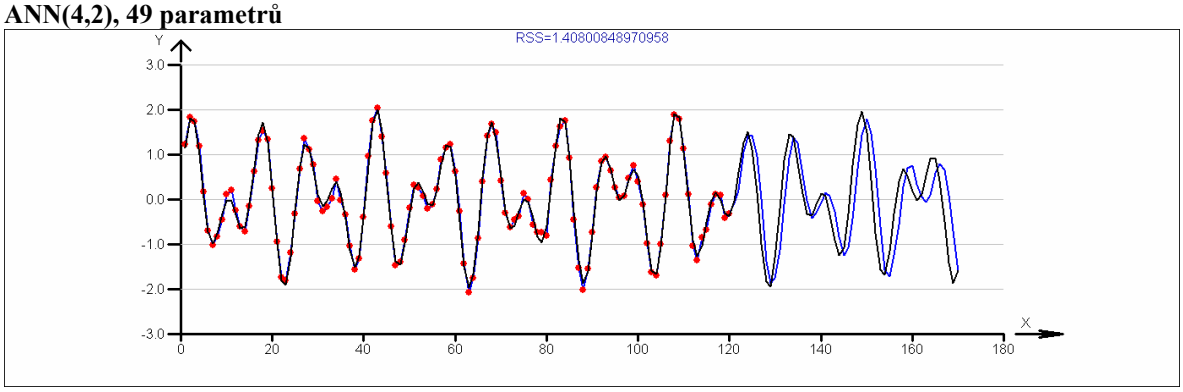
ANN(7,4), 100 parametrů



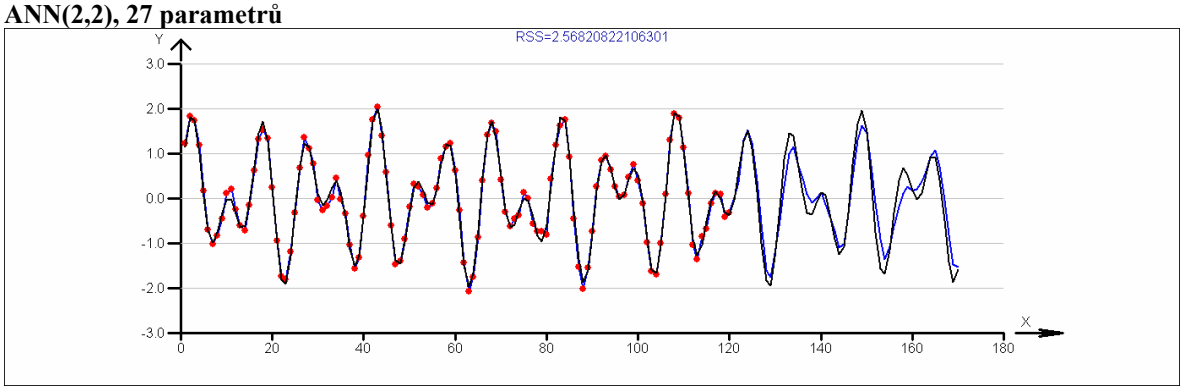
ANN(5,3), 67 parametrů



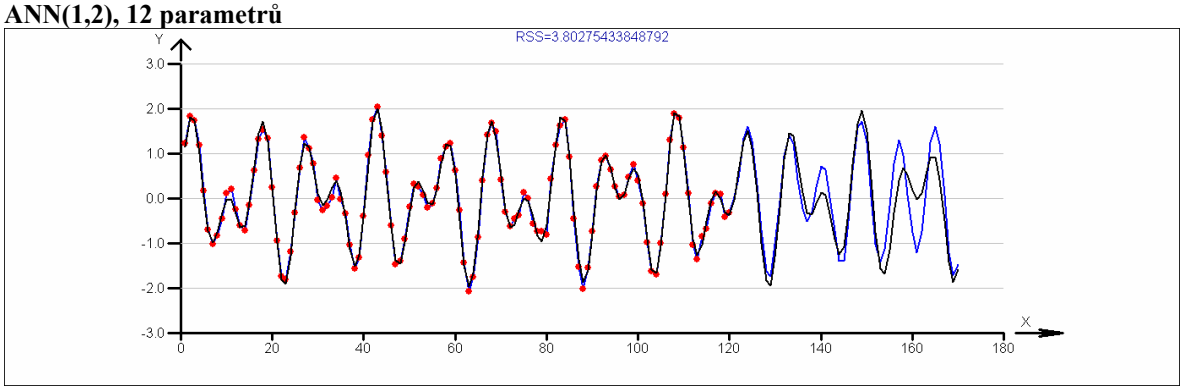
ANN(4,2), 49 parametrů



ANN(2,2), 27 parametrů



ANN(1,2), 12 parametrů



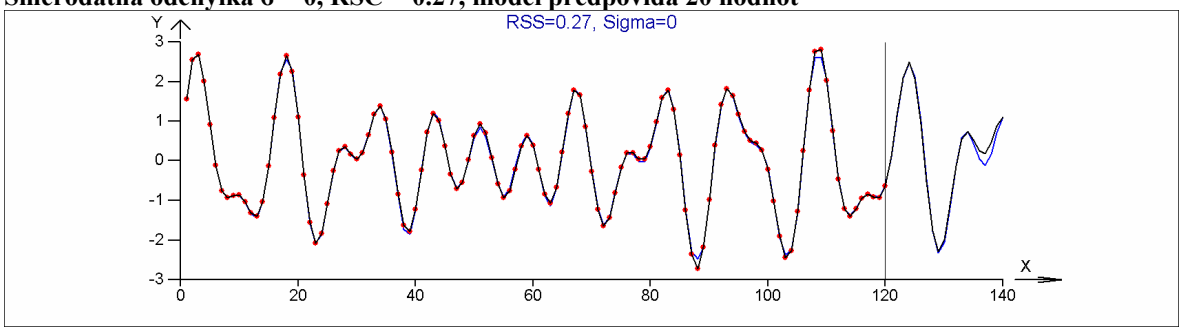
6.4. Vliv směrodatné odchylky na předpověď

Simulační studie v tomto odstavci se empiricky zabývá závislosti spolehlivosti předpovědi ANN-TS na rozptylu šumu v datech. Použitý model ani architektura ANN se nemění, je použit model ANN 9,(5,3) a tříparametrický model 2. Výsledky simulace ukazují stabilitu modelu ANN-TS i při značně velkém rozptylu dat. Použité modely a algoritmy jsou uvedeny u jednotlivých příkladů.

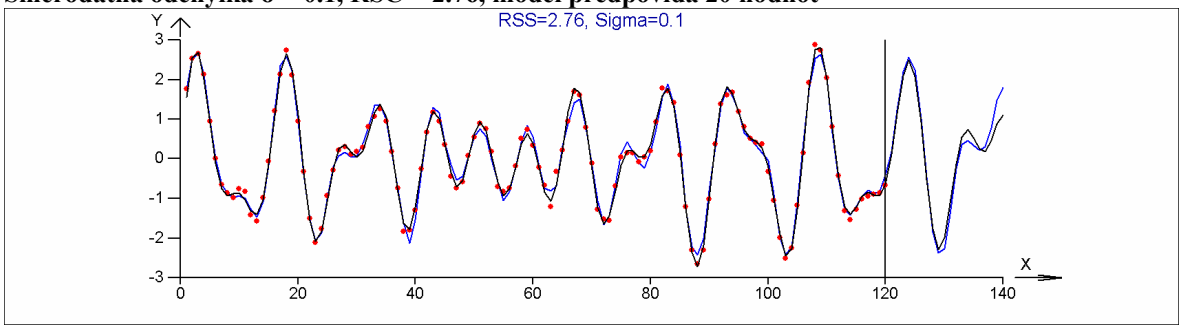
```
n=120                  // Délka simulované řady
npred=20               // Počet předpovídaných hodnot
ndpth=9                // Hloubka modelu
narch=vec(5,3)         // Architektura ANN
```

```
rnoise=0.              // Směrodatná odchylka šumu
p1=1.3;p2=2.1;p3=2.4   // Hodnoty parametrů pro generaci funkčních hodnot
xi=1:n                 // Časová základna

x=sin(xi/p1)+sin(xi/p2)+sin(xi/p3)+normalr(n)*rnoise // Model 2
```

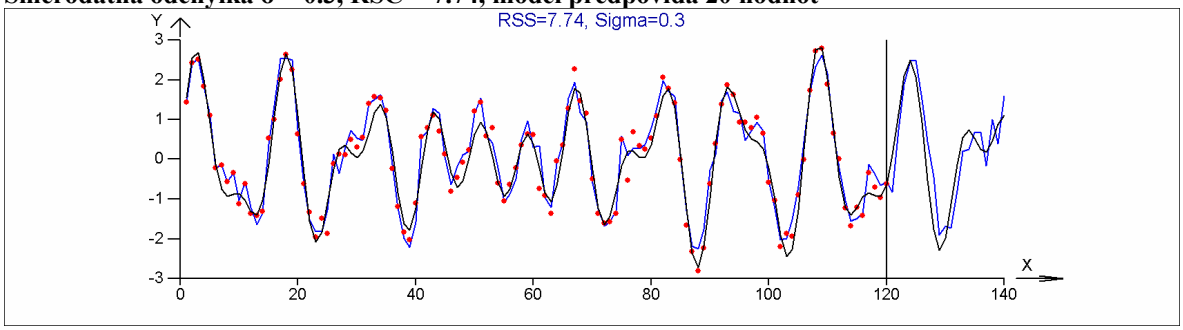
Směrodatná odchylka σ = 0, RSČ = 0.27, model předpovídá 20 hodnot



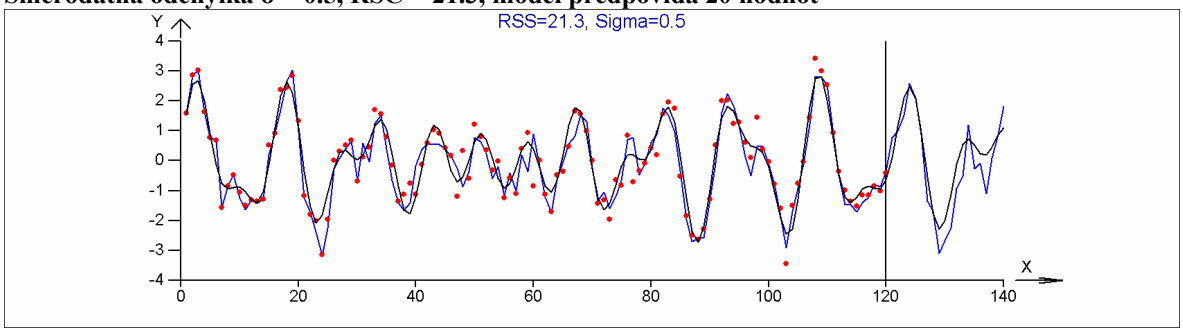
Směrodatná odchylka σ = 0.1, RSČ = 2.76, model předpovídá 20 hodnot



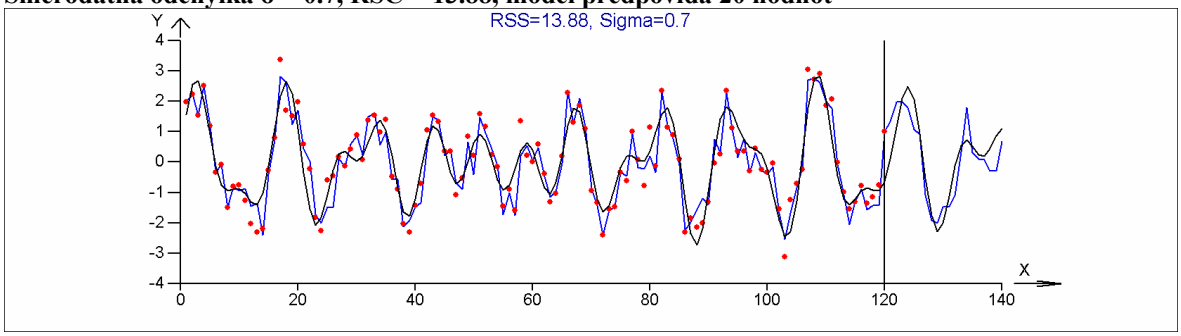
Směrodatná odchylka σ = 0.3, RSČ = 7.74, model předpovídá 20 hodnot



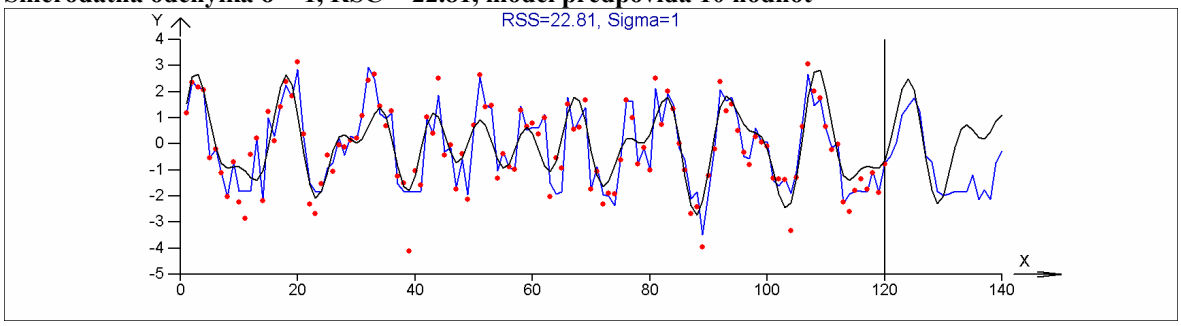
Směrodatná odchylka σ = 0.5, RSČ = 21.3, model předpovídá 20 hodnot



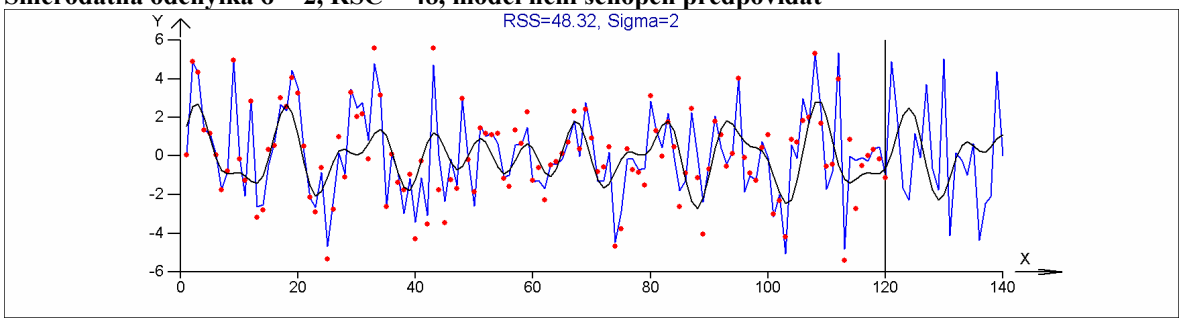
Směrodatná odchylka σ = 0.7, RSČ = 13.88, model předpovídá 20 hodnot



Směrodatná odchylka σ = 1, RSČ = 22.81, model předpovídá 10 hodnot



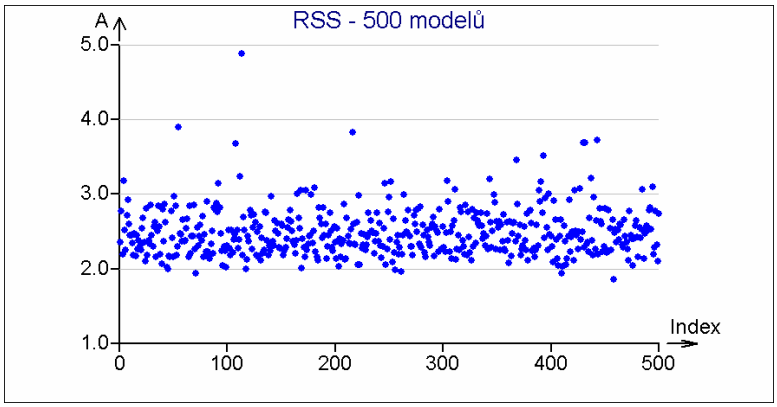
Směrodatná odchylka σ = 2, RSČ = 48, model není schopen předpovídat



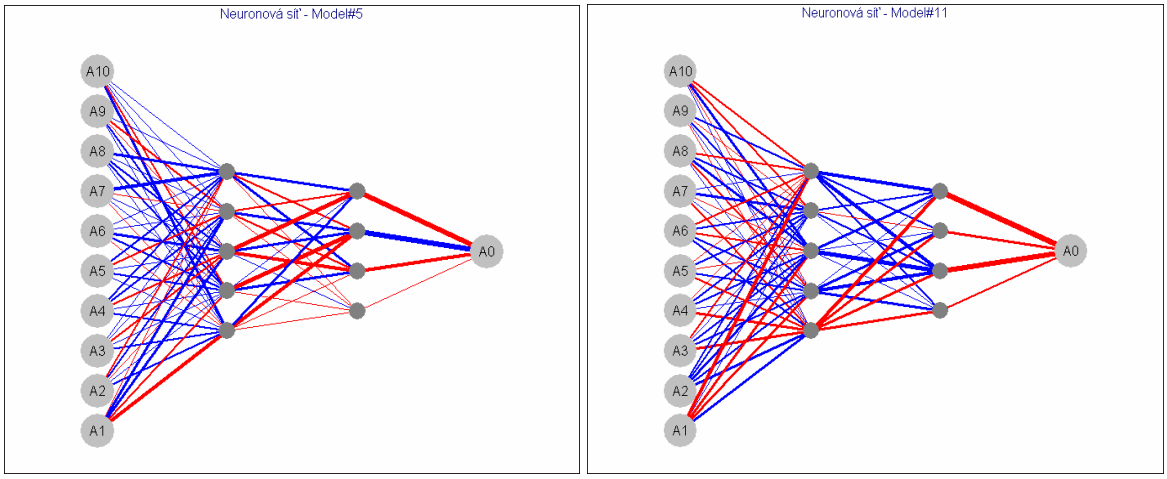
6.5. Konstrukce konfidenčního intervalu modelu

Jak bylo naznačeno v předchozích odstavcích, nestabilita hodnot parametrů modelu ANN-TS nemusí implikovat nestabilitu samotného modelu. Parametry se optimalizují derivačními algoritmy z náhodně generovaných počátečních hodnot a optimalizační algoritmus najde hodnoty parametrů, které jsou při každém výpočtu zcela různé, avšak modely jsou vyhovující, i když vždy mírně odlišné. Tohoto faktu bylo využito k empirické konstrukci Monte Carlo konfidenčních intervalů predikce a předpovědi zvoleného modelu. Pro daná data bylo konstruováno několik desítek až několik set modelů ANN-TS zvolené architektury. Kvalita proložení byla kontrolována pomocí dosaženého reziduálního součtu čtverců, viz Obr. 92. Hodnoty predikce i hodnoty předpovědi vykazují normální rozdělení. Tyto hodnoty jsou využity k odhadu střední hodnoty pomocí průměru a směrodatné odchylky, z nichž se pak získají intervaly spolehlivosti na požadované hladině významnosti. Na Obr. 96 jsou porovnány předpovědi modelu ANN-TS 8(5,3) pro deterministickou funkci sinus mírně zatíženou šumem (A), semiperiodický stochastický signál (B) a zcela

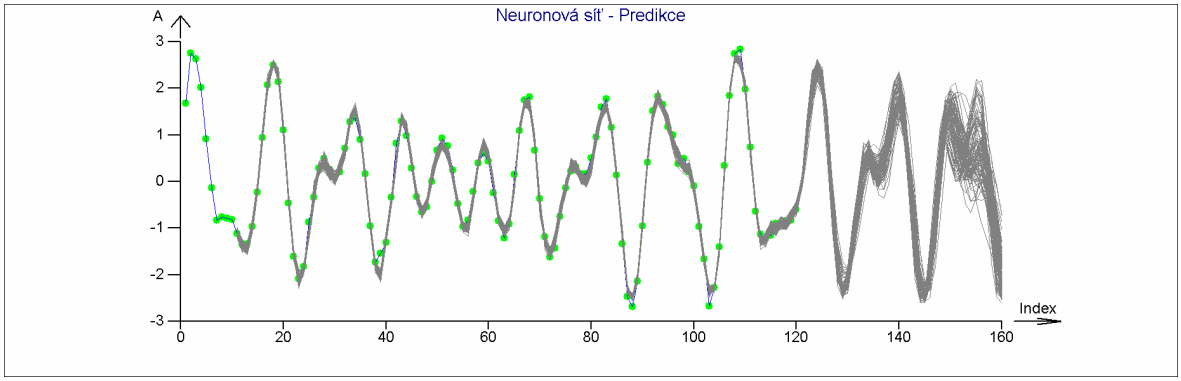
nepredikovatelný průběh Gaussovského náhodného kráčení. Předpovědi odpovídají informaci obsažené v signálech.



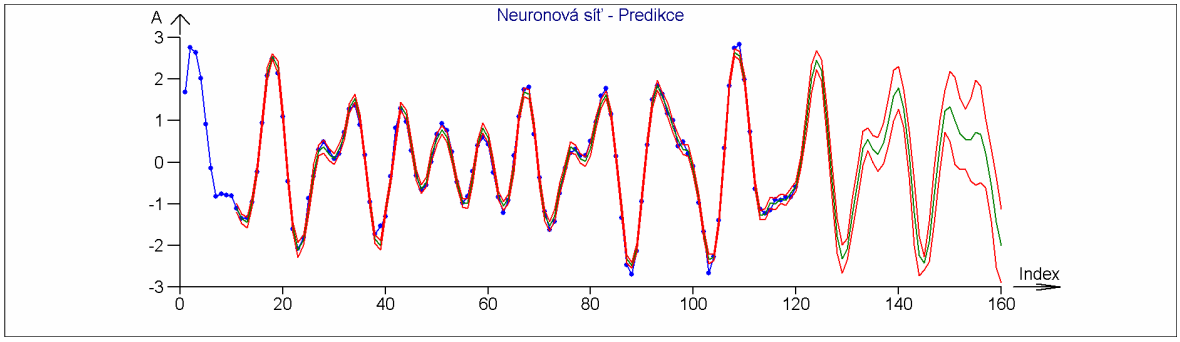
Obr. 92 Reziduální součty čtverců při simulaci 500 modelů pro stejná data leží v úzkém rozmezí (2, 3)



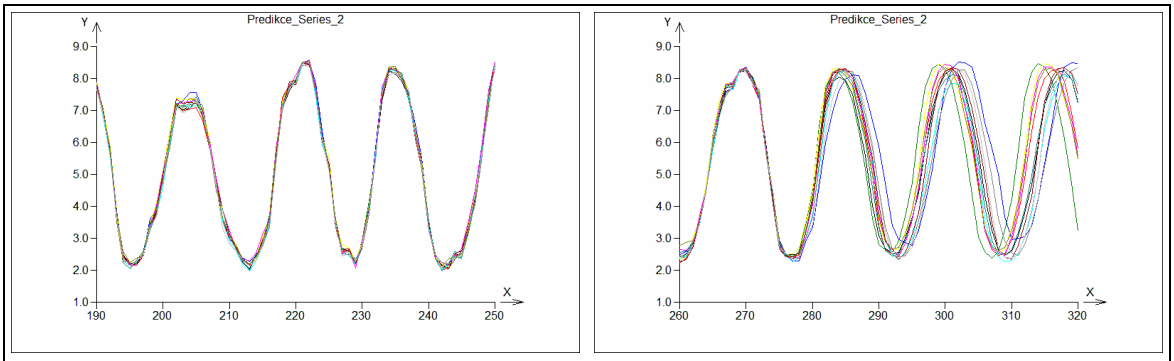
Obr. 93 Dvě neuronové sítě se stejnou strukturou ale zcela odlišnými parametry, poskytující téměř shodný průběh *x*(*t*), ale mírně odlišnou předpověď.



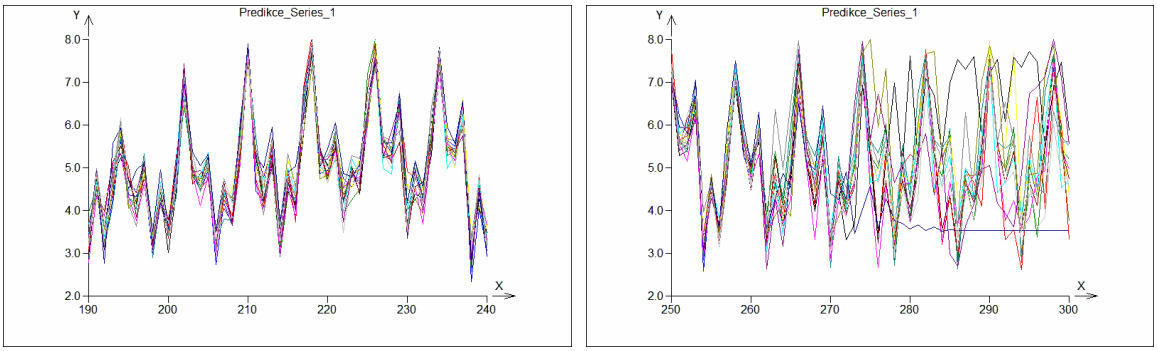
Obr. 94 Průběhy 200 modelů ANN-TS s předpovědí 40 hodnot



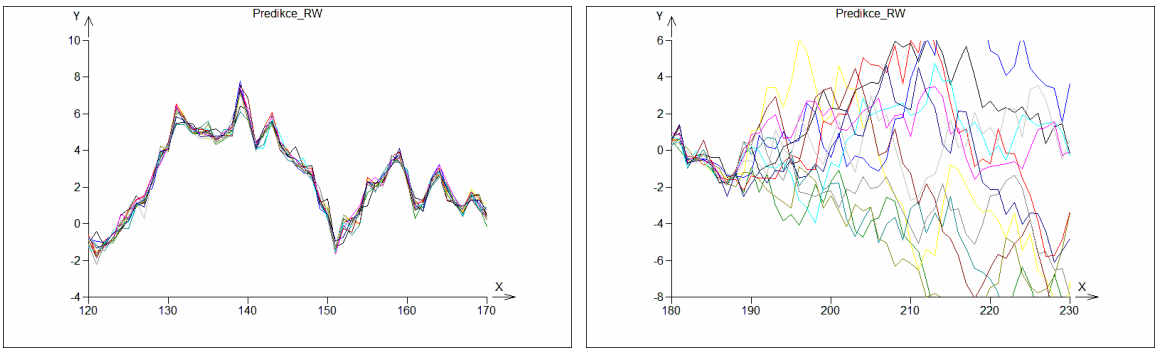
Obr. 95 Konfidenční intervaly vypočítané z průběhů 200 modelů z předchozího grafu



A. Periodický signál a jeho předpověď



B. Semiperiodický signál a jeho předpověď



C. Aperiodický náhodný signál (náhodné kráčení) a jeho předpověď

Obr. 96 Monte Carlo předpovědi pro 3 typy procesů

7. Závěr

Současné metody a výpočetní aparát aplikované statistiky poskytují značný potenciál pro vyhodnocení, modelování a interpretaci experimentálních dat ve výzkumu, technologii a kontrole kvality. Na několika úspěšných reálných studiích jsme se pokusili naznačit malý zlomek možností použití klasických, ale především neklasických, například robustních postupů statistického modelování. Statistické metody jsou v praxi zdánlivě široce používané, jsou součástí norem, SOP a smluvních dokumentací avšak využití zůstává často na úrovni „průměr, směrodatná odchylka“ a ke statistickým metodám je obecně přistupováno s nedůvěrou. V této práci ukazujeme, že využití pokročilejších metod statistické analýzy dat je možné, žádoucí a přínosné ve velmi širokém spektru oborů a aplikací.

8. Literatura

[1] Ronald A. Fisher: Statistical Methods for Research Workers (Twelfth ed.). Oliver and Boyd. ISBN 0050021702 (1954).
[2] Milan Meloun, Jiří Militký: Statistical Data Anaylsis: A Practical Guide, Woodhead Publishing (2011), ISBN 978-0857091093
[3] Tukey, J.W.: Exploratory Data Analysis. Reading, Massachusetts: Addison-Wesley Publishing Company (1977)

STATISTICKÉ ŘÍZENÍ JAKOSTI, REGULAČNÍ DIAGRAMY
[4] Shewhart, W. A. (1931) Economic Control of Quality of Manufactured Product ISBN 0-87389-076-0
[5] Shewhart, W. A. (1939) Statistical Methods from the Viewpoint of Quality Control ISBN 0-486-65232-7
[6] Deming, W. E. (1982) Out of the Crisis: Quality, Productivity and Competitive Position ISBN 0-521-30553-5
[7] Joseph Defeo and J.M. Juran: Juran's Quality Handbook: The Complete Guide to Performance Excellence 6/e (2010), McGraw-Hill, ISBN 978-0071629737
[8] Douglas C. Montgomery: Introduction to Statistical Quality Control, Wiley; 6 edition (2008), ISBN 978-0470169926
[9] Wheeler, Donald J: Understanding Variation: The Key to Managing Chaos - 2nd Edition. SPC Press, Inc. ISBN 0-945320-53-1. (1999).
[10] Adams, Cary W.; Gupta, Praveen; Charles E. Wilson (2003). Six Sigma Deployment. Burlington, MA: Butterworth-Heinemann. ISBN 0750675233.
[11] Breyfogle, Forrest W. (1999). Implementing Six Sigma: Smarter Solutions Using Statistical Methods. New York, NY: John Wiley & Sons. ISBN 0471265721.
[12] Pyzdek, Thomas and Paul A. Keller (2009). The Six Sigma Handbook, Third Edition. New York, NY, McGraw-Hill. ISBN 0071623388.
[13] Hotelling, H.: The generalization of student's ratio. Ann. Mathe. Statist., 2: 360-378 (1931).
[14] Sullivan, J.H. and W.H. Woodall: A comparison of multivariate control charts for individual observations. J. Qual. Technol., 28: 398-408 (1996).
[15] Vargas, J.A. and C.J. Lagos: Comparison of multivariate control charts for process dispersion. Qual. Eng., 19: 191-196 (2007).
[16] Tracy, N.D., J.C. Young and R.L. Mason: Multivariate control charts for individual observations. J. Qual. Technol., 24: 88-95 (1992).

CHANGE POINT
[17] Antoch J., Hušková M.: Change-point problem, Computational Aspects of Model Choice, Physica-Verlag, Heidelberg, 1993, pp. 11-38
[18] Antoch J., Hušková M., Veraverbeke N.: Change-point problem and bootstrap, J. of Nonparametric Statistics 5 (1995), 123-144
[19] Jarušková D., Detection of change point in series of river discharges, Vodohospodářský časopis 38 (1990), 501-515
[20] Hušková M.: Některé postupy pro detekci změn, Konference Analýza dat II Bohdaneč, TriloByte, podzim 2007
[21] Cobb G.W.: The problem of the Nile. Conditional solution to a change-point problem, Biometrika 65 (1978), 243-251

[22] Jaromír Antoch, Marie Huskova a Daniela Jaruskova: Change point problem po deseti letech, ROBUST'98, str. 1-42, JČMF 1998
[23] S. R. Esterby, A. H. El-Shaarawi: Inference about the Point of Change in a Regression Model, Appi. Statist. (1981), 30, No. 3, pp. 277-285
[24] J. A. Hartigan: Linear Estimators in Change Point Problems, The Annals of Statistics 1994, Vol. 22, No. 2, 824-834
[25] Barry James, Kang Ling James: Tests for a change-point, Biometrika (1987), 74, 1, pp. 71-83
[26] K. Kupka: Statistical stability and Change point detection, Invited lecture, Univ. of California Riverside, 2008

ANN
[27] Peter C. Austin and Jack V. Tu: Bootstrap Methods for Developing Predictive Models, The American Statistician, Vol. 58, No. 2 (May, 2004), pp. 131-137
[28] D. M. Wolpert, R. C. Miall: Detecting chaos with neural networks, Proceedings: Biological Sciences, Vol. 242, No. 1304 (Nov. 22, 1990), pp. 82-86
[29] J. V. Hansen and R. D. Nelson: Forecasting and Recombining Time-Series Components by Using Neural Networks, The Journal of the Operational Research Society, Vol. 54, No. 3 (Mar., 2003), pp. 307-317
[30] Tim Hill, Marcus O'Connor, William Remus: Neural Network Models for Time Series Forecasts, Management Science, Vol. 42, No. 7 (Jul. 1996), pp. 1082-1092
[31] Bing Cheng and D. M. Titterington: Neural Networks: A Review sysdbafrom a Statistical Perspective, Statistical Science, Vol. 9, No. 1 (Feb. 1994), pp. 2-30
[32] B. D. Ripley: Neural Networks and Related Methods for Classification, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 56, No. 3(1994), pp. 409-456
[33] J. T. Gene Hwang and A. Adam Ding: Prediction Intervals for Artificial Neural Networks, Journal of the American Statistical Association, Vol. 92, No. 438 (Jun., 1997), pp. 748-757
[34] Tze Leung Lai and Samuel Po-Shing Wong: Stochastic Neural Networks with Applications to Nonlinear Time Series, Journal of the American Statistical Association, Vol. 96, No. 455 (Sep., 2001), pp. 968-981
[35] Brad Warner and Manavendra Misra: Understanding Neural Networks as Statistical Tools, The American Statistician, Vol. 50, No. 4 (Nov. 1996), pp. 284-293
[36] Robert A. Kilmer, Alice E. Smith, Larry J. Shuman: Computing confidence intervals for stochastic simulation using neural network metamodels, Computers & Industrial Engineering 36 (1999) 391-407
[37] I. Rivals, L. Personnaz: Construction of confidence intervals for neural networks based on least squares estimation, Neural Networks 13 (2000) 463–484
[38] Francesco Giordano, Michele La Rocca*, Cira Perna: Forecasting nonlinear time series with neural network sieve bootstrap, Computational Statistics & Data Analysis 51 (2007) 3871–3884
[39] Ehsan Mazloumi, Geoff Rose, Graham Currie, Sara Moridpour: Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction, Engineering Applications of Artificial Intelligence 24 (2011) 534–542

ANNTS
[40] W.K. Wonga, Min Xia, W.C. Chu: Adaptive neural network model for time-series forecasting, European Journal of Operational Research 207 (2010) 807–816

[41] Feng Lin, Xing Huo Yu, Shirley Gregor and Richard Irons: Time series forecasting with neural networks, Complexity International, Volume 02 April 1995, ISSN 1320-0682
[42] Roselina Sallehuddin and Siti Mariyam Hj. Shamsuddin: Hybrid grey relational artificial neural network and auto regressive integrated moving average model for forecasting time-series data, Applied Artificial Intelligence, 23:443–486
[43] Julian Faraway: Time series forecasting with neural networks: a comparative study using the airline data, Appl. Statist. (1998) 47, part 2, pp. 231-250
[44] G. Bandyopadhyay, S. Chattopadhyay: Single hidden layer artificial neural network models versus multiple linear regression model in forecasting the time series of total ozone, Int. J. Environ. Sci. Tech., 4 (1): 141-149, 2007
[45] Edwards T, Tansley D S W, Davey N, Frank R J: Traffic Trends Analysis using Neural Networks, Proceedings of the International Workshop on Applications of Neural Networks to Telecommuncations (1997) 3. pp. 157-164
[46] Dorffner G: Neural Networks for Time Series Processing. Neural Network World (1996) 4/96, 447-468

ANNTS, APLIKACE
[47] Gwo-Fong Lin, Ming-Chang Wu: A hybrid neural network model for typhoon-rainfall forecasting, Journal of Hydrology 375 (2009) 450–458
[48] M. Caselli, L. Trizio, G. de Gennaro, P. Ielpo: A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model, Water Air Soil Pollut (2009) 201:365–377
[49] Hafzullah Aksoy Æ Ahmad Dahamsheh: Artificial neural network models for forecasting monthly precipitation in Jordan, Stoch Environ Res Risk Assess (2009) 23:917–931
[50] Siti M. Shamsuddin, Roselina Sallehuddin and Norfadzila M. Yusof: Artificial Neural Network Time Series Modeling for Revenue Forecasting, Chiang Mai J. Sci. 2008; 35(3) : 411-426
[51] Veysel Güldal, Hakan Tongal: Comparison of Recurrent Neural Network, Adaptive Neuro-Fuzzy Inference System and Stochastic Models in Egirdir Lake Level Forecasting, Water Resour Manage (2010) 24:105–128
[52] Chang-Shian Chen, Boris Po-Tsang Chen, Frederick Nai-Fang Chou, Chao-Chung Yang: Development and application of a decision group Back-Propagation Neural Network for flood forecasting, Journal of Hydrology 385 (2010) 173–182
[53] Whei-Min Lin a, Hong-Jey Gowa, Ming-Tang Tsai: Electricity price forecasting using Enhanced Probability Neural Network, Energy Conversion and Management 51 (2010) 2707–2714
[54] Chan Man-Chung, Wong Chi-Cheong, Lam Chi-Chung: Financial Time Series Forecasting by Neural Network Using Conjugate Gradient Learning Algorithm and Multiple Linear Regression Weight Initialization, Department of Computing The Hong Kong Polytechnic University Kowloon, Hong Kong
[55] Lean Yu, Shouyang Wang, Kin Keung Lai: Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm, Energy Economics 30 (2008) 2623–2635
[56] Gorka Landeras, Amaia Ortiz-Barredo, José Javier López: Forecasting Weekly Evapotranspiration with ARIMA and Artificial Neural Network Models, Journal of irrigation and drainage engineering, asce, May/June 2009

[57] Lee, TS; Chiu, CC: Neural network forecasting of an opening cash price index, International Journal of Systems Science, 33 (3): 229-237 FEB 20 2002
[58] Chen HB, Grant-Muller S, Mussone L, Montgomery F: A study of hybrid neural network approaches and the effects of missing data on traffic forecasting, Neural Computing & Applications, 10 (3): 277-286 2001
[59] Hall T, Brooks HE, Doswell CA: Precipitation forecasting using a neural network, Weather and Forecasting, 14 (3): 338-345 JUN 1999
[60] Tsai CP, Lee TL: Back-propagation neural network in tidal-level forecasting, Journal of Waterway Port Coastal and Ocean Engineering-ASCE, 125 (4), 195-202 JUL-AUG 1999
[61] S. SriLakshmi, R.K.Tiwari: Model dissection from earthquake time series: A comparative analysis using modern non-linear forecasting and artificial neural network approaches, Computers & Geosciences 35 (2009) 191–204

SMĚS ROZDĚLENÍ

[62] Keon -Tae Sohn and Jee -Seon Baik: Estimation in a mixture normal distribution, Journal of Applied Mathematics and Computing, Volume 4, Number 1, 223-233
[63] Rahman, Mezbahur; Rahman, Rumanur; Pearson, Larry M.: Quantiles for Finite Mixtures of Normal Distributions, International Journal of Mathematical Education in Science & Technology, v37 n3 p. 352-358 Apr 2006
[64] A. Durio E.D. Isaia: A quick procedure for model selection in the case of mixture of normal densities, Computational Statistics & Data Analysis 51 (2007) 5635 – 5643
[65] Erik Meijer: A Simple Identification Proof for a Mixture of Two Univariate Normal Distributions, Journal of Classification 25:113-123 (2008)
[66] Xin Liu, Yongzhao Shao: Asymptotics for the likelihood ratio test in a two-component normal mixture model, Journal of Statistical Planning and Inference 123 (2004) 61 – 81
[67] D. P. Vetrov, D. A. Kropotov, and A. A. Osokin: Automatic Determination of the Number of omponents in the EM Algorithm of Restoration of a Mixture of Normal Distributions, Computational Mathematics and Mathematical Physics, 2010, Vol. 50, No. 4, pp. 733–746
[68] Jin Wang, Chunlei Liu: Generating multivariate mixture of normal distributions using a modified Cholesky decomposition, Proceedings of the 2006 Winter Simulation Conference
[69] K. E. Basford and G. J. McLachlan: Likelihood Estimation with Normal Mixture Models, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 34, No. 3(1985), pp. 282-289
[70] Tony S. Wirjanto and Dinghai Xu: The Applications of Mixtures of Normal Distributions in Empirical Finance: A Selected Survey
[71] Constantinos Petropoulos: Estimation of a quantile in a mixture model of exponential distributions with unknown location and scale parameter, Bulletin of University of the Aegean, Samos, Greece
[72] Hironori Fujisawa, Shinto Eguchi: Robust estimation in the normal mixture model, Journal of Statistical Planning and Inference 136 (2006) 3989 – 4011
[73] Quandt, R. E. and J. B. Ramsey, Estimating Mixtures of Normal Distributions and Switching Regressions, Journal of the American Statistical Association, 73, 730-738 (1978)
[74] McLachlan, G., On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture, Applied Statistics, 38: 318 – 324 (1987)

LP NORMA
[75] P. Y. Lai and Stephen M. S. Lee: An Overview of Asymptotic Properties of Lp Regression under General Classes of Error Distributions, Journal of the American Statistical Association, Vol. 100, No. 470 (Jun., 2005), pp. 446-458
[76] Hideki Kosaki: Applications of Uniform Convexity of Noncommutative Lp-Spaces, Transactions of the American Mathematical Society, Vol. 283, No. 1 (May, 1984), pp. 265-282
[77] D. Song and A. K. Gupta: Lp-Norm Uniform Distribution, Proceedings of the American Mathematical Society, Vol. 125, No. 2 (Feb., 1997), pp. 595-601
[78] Lawrence G. Brown and Bradley J. Lucier: Best Approximations in L1 are Near Best in Lp, p<1, Proceedings of the American Mathematical Society, Vol. 120, No. 1 (Jan., 1994), pp. 97-100
[79] Lajos Horvath: On Lp-Norms of Multivariate Density Estimators, The Annals of Statistics, Vol. 19, No. 4 (Dec., 1991), pp. 1933-1949
[80] Pinkus, Alan: On L1 approximation, Cambridge University Press 1989, ISBN 0-521-36650-X
[81] A. N. Tyrsin: Robust construction of regression models based on the generalized least absolute deviations method, Journal of Mathematical Sciences, Volume 139, Number 3, 6634-6642

SPLINE, PIECEWISE REGRESE
[82] Blischke, W. R., "Least squares estimates of two intersecting lines," Technical Report No, 7, Department of Navy, ONR, Cornell University (1961).
[83] R. E. Quandt: The estimation of the parameters of a linear regression system obeying two separate regimes, J Am Statist. Ass., 53, 873-880.
[84] Judith D. Toms and Mary L. Lesperance: Piecewise Regression: A Tool for Identifying Ecological Thresholds, Ecology, Vol. 84, No. 8 (Aug., 2003), pp. 2034-2041
[85] Asher Tishler and Israel Zang: A New Maximum Likelihood Algorithm for Piecewise Regression, Journal of the American Statistical Association, Vol. 76, No. 376 (Dec., 1981), pp. 980-987
[86] Adil Bagirov, Conny Clausen, Michael Kohler: An algorithm for the estimation of a regression function by continuous piecewise linear functions, Comput Optim Appl (2010) 45: 159–179
[87] B. H. Rosman: Another Approach to the Cubic Interpolating Spline, The American Mathematical Monthly, Vol. 80, No. 8 (Oct., 1973), pp. 927-930
[88] Steven A. Julious: Inference and estimation in a changepoint regression problem, The Statistician (2001), 50, Part 1, pp. 51-61
[89] Krisnaiah, P. K. and Miao, B. Q. (1988) Review about estimation of change-points. In Handbook of Statistics (eds P. K. Krisnaiah and C. R. Rao), vol. 7, pp. 375-402. Amsterdam: North-Holland
[90] Hudson, D. J. (1966) Fitting segmented curves whose join points have to be estimated. J Am. Statist. Ass., 61, 1097-1129
[91] Shaban, S. A. (1980). Change point problem and two-phase regression: An annotated bibliography. Int. Statist. Rev. 48, 83-93.
[92] Wei Biao Wu, Michael Woodroofe, Graciela Mentz: Isotonic Regression: Another Look at the Changepoint Problem, Biometrika, Vol. 88, No. 3 (Sep., 2001), pp. 793-804
[93] Jianhua Z. Huang: Local Asymptotics for Polynomial Spline Regression, The Annals of Statistics, Vol. 31, No. 5 (Oct., 2003), pp. 1600-1635

[94] Robison, D. E., "Estimates for the Points of Intersection of Two Polynomial Regressions," Journal of the American Statistical Association, 59 (March 1964), 214-240
[95] Victor E. McGee and Willard T. Carleton: Piecewise Regression, Journal of the American Statistical Association, Vol. 65, No. 331 (Sep. 1970), pp. 1109-1124
[96] Dale J. Poirier: Piecewise Regression Using Cubic Spline, Journal of the American Statistical Association, Vol. 68, No. 343 (Sep., 1973), pp. 515-524
[97] Gihan F. Malash, Mohammad I. El-Khaiary: Piecewise linear regression: A statistical method for the analysis of experimental adsorption data by the intraparticle-diffusion models, Chemical Engineering Journal 163 (2010) 256–263
[98] Douglas M. Hawkins: Point Estimation of the Parameters of Piecewise Regression Models, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 25, No. 1 (1976), pp. 51-57
[99] Hyune-Ju Kim, Michael P. Fay, Binbing Yu, Michael J. Barrett, Eric J. Feuer: Comparability of Segmented Line Regression Models, Biometrics, Vol. 60, No. 4 (Dec., 2004), pp. 1005-1014
[100] Richard Bellman and Robert Roth: Curve Fitting by Segmented Straight Lines, Journal of the American Statistical Association, Vol. 64, No. 327 (Sep., 1969), pp. 1079-1084
[101] D. E. Robison: Estimates for the Points of Intersection of Two Polynomial Regressions: Journal of the American Statistical Association, Vol. 59, No. 305 (Mar., 1964), pp. 214-224
[102] David W. Bacon and Donald G. Watts: Estimating the Transition between Two Intersecting Straight Lines, Biometrika, Vol. 58, No. 3 (Dec. 1971), pp. 525-534
[103] Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. Biometrika 56, 495-504.
[104] David W. Bacon and Donald G. Watts: Estimating the Transition between Two Intersecting Straight Lines, Biometrika, Vol. 58, No. 3 (Dec. 1971), pp. 525-534
[105] A. R. Gallant and Wayne A. Fuller: Fitting Segmented Polynomial Regression Models Whose Join Points have to be Estimated, Journal of the American Statistical Association, Vol. 68, No. 341 (Mar., 1973), pp. 144-147 (UKAZUJE, ZE TO JDE)
[106] Hartley, H.O.: The Modified Gauss-Newton Method for the Fitting of Non-Linear Regression Functions by Least Squares, Technometrics, 3 (May 1961), 269-80. (DOKAZUJE, ZE TO NEJDE.)
[107] Bates, D. M. and Watts, D. G.: Nonlinear Regression Analysis and Its Applications. New York: Wiley (1988). (NEKDY TO NEJDE.)
[108] A. Ronald Gallant: Testing a Nonlinear Regression Specification: A Nonregular Case, Journal of the American Statistical Association, Vol. 72, No. 359 (Sep., 1977), pp. 523-530
[109] David O. Siegmund, Heping Zhang: Confidence regions in broken line regression, Change point problems in IMS Lecture notes - Monograph Series Vol. 23 (1994), Stanford and Yale University
[110] T. Goto et al.: A Robust Spline Filter on the basis of L2-norm, Precision Engineering 29 (2005) 157–161
[111] Chunming Zhang: Assessing the equivalence of nonparametric regression tests based on spline and local polynomial smoothers, Journal of Statistical Planning and Inference 126 (2004) 73 – 95
[112] Wen Hsiang Wei: Derivatives diagnostics and robustness for smoothing splines, Computational Statistics & Data Analysis 46 (2004) 335 – 356
[113] Peide Shi: M-type regression splines involving time series, Journal of Statistical Planning and Inference 61 (1997) 17-37

[114] Lin-An Chen: Multivariate regression splines, Computational Statistics & Data Analysis 26 (1997) 71-82
[115] Zulfiqar Habib, Muhammad Sarfraz, Manabu Sakai: Rational cubic spline interpolation with shape control, Computers & Graphics 29 (2005) 594–605
[116] Igor Averbakha, Yun-Bin Zhao: Robust univariate spline models for interpolating interval data, Operations Research Letters 39 (2011) 62–66
[117] Hao Cheng, Shu-Cherng Fanga, John E. Lavery: Shape-preserving properties of univariate cubic L1 splines, Journal of Computational and Applied Mathematics 174 (2005) 361–382

SPLINES, LOKÁLNÍ REGRESE

[118] Stefan Klanke, Sethu Vijayakumar, Stefan Schaal: A Library for Locally Weighted Projection Regression, Journal of Machine Learning Research 9 (2008) 623-626
[119] Wolfgang Hardle and Adrian W. Bowman: Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands, Journal of the American Statistical Association, Vol. 83, No. 401 (Mar., 1988), pp. 102-110
[120] Gerda Claeskens and Ingrid Van Keilegom: Bootstrap Confidence Bands for Regression Curves and Their Derivatives, The Annals of Statistics, Vol. 31, No. 6 (Dec., 2003), pp. 1852-1884
[121] William S. Krasker and Roy E. Welsch: Efficient Bounded-Influence Regression Estimation, Journal of the American Statistical Association, Vol. 77, No. 379 (Sep., 1982), pp. 595-604
[122] S. Zhou, X. Shen, D. A. Wolfe: Local Asymptotics for Regression Splines and Confidence Regions, The Annals of Statistics, Vol. 26, No. 5 (Oct., 1998), pp. 1760-1782
[123] Keming Yu and M. C. Jones: Local Linear Quantile Regression, Journal of the American Statistical Association, Vol. 93, No. 441 (Mar. 1998), pp. 228-237
[124] F. Jay Breidt and Jean D. Opsomer: Local Polynomial Regression Estimators in Survey Sampling, The Annals of Statistics, Vol. 28, No. 4 (Aug., 2000), pp. 1026-1053
[125] Rafael A. Irizarry: Local Regression with Meaningful Parameters, The American Statistician, Vol. 55, No. 1 (Feb., 2001), pp. 72-79
[126] William S. Cleveland and Susan J. Devlin: Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, Journal of the American Statistical Association, Vol. 83, No. 403 (Sep., 1988), pp. 596-610
[127] D. Ruppert and M. P. Wand: Multivariate Locally Weighted Least Squares Regression, The Annals of Statistics, Vol. 22, No. 3 (Sep., 1994), pp. 1346-1370
[128] William S. Cleveland: Robust Locally Weighted Regression and Smoothing Scatterplots, Journal of the American Statistical Association, Vol. 74, No. 368 (Dec., 1979), pp. 829-836
[129] Ferdinand T. Wang and David W. Scott: The L1 Method for Robust Nonparametric Regression, Journal of the American Statistical Association, Vol. 89, No. 425 (Mar., 1994), pp. 65-76
[130] Hans-Georg Muller: Weighted Local Regression and Kernel Methods for Nonparametric Curve Fitting, Journal of the American Statistical Association, Vol. 82, No. 397 (Mar., 1987), pp. 231-238
[131] Jaromír Antoch, Gerard Gregoire, Marie Hušková: Tests for continuity of regression functions, Journal of Statistical Planning and Inference 137 (2007) 753 – 777

REGRESE
[132] Draper, N.R. and Smith, H.: Applied Regression Analysis Wiley Series in Probability and Statistics (1998)
[133] Birkes, David and Dodge, Y.: Alternative Methods of Regression. ISBN 0-471-56881-3
[134] David A. Ratkowsky: Nonlinear Regression Modeling: A Unified Practical Approach, Marcel Dekker Inc (November 13, 1989), ISBN 978-0824781897
[135] Leo Breiman: Statistical Modeling: The Two Cultures, Statistical Science, Vol. 16, No. 3 (Aug., 2001), pp. 199-215
[136] Jinyan Fan, Jianyu Pan: A note on the Levenberg–Marquardt parameter, Applied Mathematics and Computation 207 (2009) 351–359
[137] Chong Li, Wen-Hong Hhang, Xiao-Qing Jin: Convergence and Uniqueness Properties of Gauss-Newton's Method, Computers and Mathematics with Applications 47 (2004) 1057-1067
[138] Changfeng Ma, Lihua Jiang, Desheng Wang: The convergence of a smoothing damped Gauss–Newton method for nonlinear complementarity problem, Nonlinear Analysis: Real World Applications 10 (2009) 2072–2087

ROBUSTNÍ REGRESE
[139] Holland, P. W., & Welsch, R. E. Robust regression using iteratively reweighted least-squares. Communications in Statistics, 1977, A6, 813-827.
[140] Andrews, D. F., & Pregibon, D. Finding the outliers that matter, Journal of the Royal Statistical Society Part C Applied Statistics, 1978, 27, 85-93.
[141] Belsley, D. A., Kuh, E., & Welsch, R. E. Regression diagnostics: Identifying influential data and sources of collinearity. New York: Wiley, 1980.
[142] Hoaglin, D. C., & Welsch, R. E. The hat matrix in regression and ANOVA. The American Statistician, 1978, 32, 17-22 and Corrigenda, 1978, 32, 146.
[143] Krasker, W. S., & Welsch, R. E. Efficient bounded-influence regression estimation using alternative definitions of sensitivity. Technical Report#3, M.I.T. Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1979.
[144] Friedrich-Wilhelm Scholz: Weighted Median Regression Estimates, The Annals of Statistics, Vol. 6, No. 3 (May, 1978), pp. 603-609
[145] Alan D. Chave and David J. Thomson: A Bounded Influence Regression Estimator Based on the Statistics of the Hat Matrix, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 52, No. 3(2003), pp. 307-322
[146] Raymond J. Carroll and A. H. Welsh: A Note on Asymmetry and Robustness in Linear Regression, The American Statistician, Vol. 42, No. 4 (Nov., 1988), pp. 285-287
[147] Javier González, Daniel Pena and Rosario Romera: A robust partial least squares regression method with applications, J. Chemometrics 2009, 23: 78–90
[148] I. N. Wakeling, H. J. H. Macfie: A robust PLS procedure, Journal Of Chemometrics, VOL. 6, 189-198 (1992)
[149] Frank Hampel, Christian Hennig, Elvezio Ronchetti: A smoothing principle for the Huber and other location M-estimators, Computational Statistics and Data Analysis 55 (2011) 324-337
[150] James E. Mays, Jeffrey B. Birch, Richard L Einsporn: An overview of model-robust regression, J. Statist. Comput. Simul. 2000, Vol.66, pp. 79-100

[151] Douglas M. Hawkins, David Olive: Applications and algorithms for least trimmed sum of absolute deviations regression, Computational Statistics & Data Analysis 32 (1999) 119-134
[152] J. D. Naranjo and T. P. Hettmansperger: Bounded Influence Rank Regression, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 56, No. 1(1994), pp. 209-220
[153] Arnold J. Stromberg and David Ruppert: Breakdown in Nonlinear Regression, Journal of the American Statistical Association, Vol. 87, No. 420 (Dec., 1992), pp. 991-997
[154] William S. Krasker and Roy E. Welsch: Efficient Bounded-Influence Regression Estimation, Journal of the American Statistical Association, Vol. 77, No. 379 (Sep., 1982), pp. 595-604
[155] Ola Hoessjer: Exact computation of the least trimmed squares estimate in simple linear regression, Computational Statistics & Data Analysis 19 (1995) 265-282
[156] A. C. Atkinson: Fast Very Robust Methods for the Detection of Multiple Outliers, Journal of the American Statistical Association, Vol. 89, No. 428 (Dec., 1994), pp. 1329-1339
[157] Frank R. Hampel: The Influence Curve and Its Role in Robust Estimation, Journal of the American Statistical Association, Vol. 69, No. 346 (Jun., 1974), pp. 383-393
[158] Frank R. Hampel: What can the foundations discussion contribute to data analysis? And what may be some of the future directions in robust methods and data analysis?, Journal of Statistical Planning and Inference 57 (1997) 7-19
[159] Chen-Chia Chuanga, Zne-Jung Leeb: Hybrid robust support vector machines for regression with outliers, Applied Soft Computing 11 (2011) 64–72
[160] Terry E. Dielman: Least absolute value regression: recent contributions, Journal of Statistical Computation and Simulation Vol. 75, No. 4. April 2005. 263-286
[161] Peter J. Rousseeuw: Least Median of Squares Regression, Journal of the American Statistical Association, Vol. 79, No. 388 (Dec., 1984), pp. 871-880
[162] A. Giloni, M. Padberg:Least Trimmed Squares Regression, Least Median Squares Regression, and Mathematical Programming: Mathematical and Computer Modelling 35 (2002) 1043-1060
[163] Peide Shi: M-type regression splines involving time series, Journal of Statistical Planning and and inference 61 (1997) 17-37
[164] Kang-Mo Jung:Multivariate least-trimmed squares regression estimator,Computational Statistics & Data Analysis 48 (2005) 307 – 316
[165] José Agulló: New algorithms for computing the least trimmed squares regression estimator, Computational Statistics & Data Analysis 36 (2001) 425– 439
[166] Charles J. Stone: Nonparametric M-regression with free knot splines, Journal of Statistical Planning and Inference 130 (2005) 183 – 206
[167] Robert Hable, Andreas Christmann: On qualitative robustness of support vector machines, Journal of Multivariate Analysis 102 (2011) 993–1007
[168] Juan A. Gil And Rosario Romera: On robust partial least squares (PLS) methods: J. Chemometrics 12, 365–378 (1998)
[169] Víctor J. Yohai a, 1, Ruben H. Zamar: Optimal locally robust M-estimates of regression, Joumal of Statistical Planning and Inference 64 (1997) 309-323
[170] Bettina Liebmann, Peter Filzmoser, Kurt Varmuza: Robust and classical PLS regression compared, J. Chemometrics 2010; 24: 111–120
[171] Lei Huang, Bai-Ling Zhang, Qian Huang: Robust interval regression analysis using neural networks, Fuzzy Sets and Systems 97 (1998) 337-347

[172] M.Hubert and K.Vanden Branden: Robust methods for partial least squares regression, J. Chemometrics 2003; 17: 537–549
[173] Uwe Kruger, Yan Zhouay, Xun Wang, David Rooney, Jillian Thompson: Robust partial least squares regression: Part II, new algorithm and benchmark studies, J. Chemometrics 2008, 22 14–22
[174] Uwe Kruger, Yan Zhou, Xun Wang, David Rooney, Jillian Thompson: Robust partial least squares regression: Part I, algorithmic developments, J. Chemometrics 2008, 22 1–13
[175] Petr J. Rousseeuw: Robust regression, positive breakdown in, Encyclopedia of Statistical Sciences, 2006 John Wiley & Sons, Inc.
[176] Peter J. Rousseeuw and Mia Hubert: Robust statistics for outlier detection, Wire Data Mining and Knowledge Discovery, Volume 1, January/February 2011
[177] K. Vanden Branden, M. Hubert: Robustness properties of a robust partial least squares regression method, Analytica Chimica Acta 515 (2004) 229–241
[178] Stefan Van Aelst, Peter J. Rousseeuw, Mia Hubert, and Anja Struyf: The Deepest Regression Method, Journal of Multivariate Analysis 81, 138–166 (2002)
[179] Ferdinand T. Wang and David W. Scott: The L1 Method for Robust Nonparametric Regression, Journal of the American Statistical Association, Vol. 89, No. 425 (Mar., 1994), pp. 65-76
[180] Hendrik Bode, Frederick Mosteller, John Tukey, Charles Winsor: The Education of a Scientific Generalist, Science, New Series, Vol. 109, No. 2840 (Jun. 3, 1949), pp. 553-558
[181] Peter J. Rousseeuw: Tutorial to robust statistics, Journal of chemometrics, vol. 5 , 1-20 (1991)
[182] Friedrich-Wilhelm Scholz: Weighted Median Regression Estimates, The Annals of Statistics, Vol. 6, No. 3 (May, 1978), pp. 603-609
[183] Jurečková, J.: Nonparametric estimates of regression coefficients, Ann. Math. Statist. 42 (1971) 1328-1338

GENERAL ROBUSNESS:
[184] F. R. Hampel: The influence curve and its role in robust estimation, J. Am. Stat. Assoc. 69, 383-393 (1974).
[185] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel: Robust Statistics: the Approach Based on Influence Functions, Wiley, New York (1986).
[186] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers and J. W. Tukey: Robust Estimates of Location: Survey and Advances, Princeton University Press, Princeton, NJ (1972).
[187] J. L. Hodges Jr.: Efficiency in normal samples and tolerance of extreme values for some estimates of location, in Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability, Vol. 1, pp.163-168, University of California Press, Berkeley and Los Angeles, CA (1967).
[188] F. R. Hampel: A general qualitative definition of robustness, Ann. Math. Stat. 42, 1887-1896 (1971).
[189] D. L. Donoho and P. J. Huber: The notion of breakdown point, in A Festschrift for Erich Lehmann, ed. by P. Bickel, K. Doksum and J. L. Hodges Jr., pp 157-184 Wadsworth, Belmont, CA (1983).
[190] Langenberg, P. and Iglewicz, B. 1986. Trimmed mean X-bar and R charts. J. Qual. Technol., 18: 152-161
[191] Moustafa Omar Ahmed Abu-Shawiesh: A Simple Robust Control Chart Based on MAD, Journal of Mathematics and Statistics 4 (2): 102-107, 2008

[192] Shabbak, A., H. Midi and M.N. Hassan, 2011. The performance of robust multivariate statistical control charts based on different cutoff-points with sustained shift in mean. J. Applied Sci., 11: 56-65.
[193] Jensen, W.A., J.B. Birch and W.H. Woodall, 2007. High breakdown estimation methods for phase I multivariate control charts. Qual. Reliabil. Eng. Int., 23: 615-629.
[194] Rocke, D.M.: Robust control charts, Technometrics, 31(2), 173–184 (1989).
[195] Rocke, D.M.: XQ and RQ charts: robust control charts, The Statistician, 41, 97–104 (1992).
[196] David Ruppert Raymond J. Carroll: Trimmed Least Squares Estimation in the Linear Model, Journal of the American Statistical Association, Vol. 75, No. 372 (Dec., 1980), pp. 828-838
[197] Effat Moussa-Hamouda and Fred C. Leone: Efficiency of Ordinary Least Squares Estimators from Trimmed and Winsorized Samples in Linear Regression, Technometrics, Vol. 19, No. 3 (Aug., 1977), pp. 265-273
[198] Mara Tableman: The influence functions for the least trimmed squares and the least trimmed absolute deviations estimators, Statistics & Probability Letters, Volume 19, Issue 4, 15 March 1994, Pages 329-337
[199] José Agulló: New algorithms for computing the least trimmed squares regression estimator, Computational Statistics & Data Analysis, Volume 36, Issue 4, 28 June 2001, Pages 425-439
[200] Kang-Mo Jung: Multivariate least-trimmed squares regression estimator, Computational Statistics & Data Analysis, Volume 48, Issue 2, 1 February 2005, Pages 307-316
[201] A. Giloni, M. Padberg: Least trimmed squares regression, least median squares regression, and mathematical programming, Mathematical and Computer Modelling, Volume 35, Issues 9-10, May 2002, Pages 1043-1060
[202] D. Vandev: A note on the breakdown point of the least median of squares and least trimmed squares estimators, Statistics & Probability Letters, Volume 16, Issue 2, 27 January 1993, Pages 117-119
[203] Ricardo A. Maronna, Douglas R. Martin, Victor J. Yohai: Robust Statistics: Theory and Methods (Wiley Series in Probability and Statistics), Wiley; 1 edition (June 14, 2006), ISBN 0470010924
[204] Tukey, J.W., 1960. A Survey of Sampling from Contaminated Distributions. In: Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling, Olkin, I., et al., (Eds.). Stanford: Stanford University Press. pp: 448-485.
[205] Rousseeuw, P.J. and C. Croux, 1993. Alternatives to the median absolute deviation. J. Am. Stat. Assoc., 80: 1273-1283.

OSTATNÍ
[206] M. Akrout et al.: Numerical and Experimental Study of the Erichsen Test for Metal Stampling, Advances in Production Engineering and Management, 3 (2008) 2, 81-92
[207] ISO 8490:1986: Metallic materials; Sheet and strip; Modified Erichsen cupping test, International Organization for Standardization, 01-Oct-1986
[208] ČSN ISO 8490 (420406), Dat.vydání: 1.9.1994, Kovové materiály. Plechy a pásy. Modifikovaná zkouška hloubením podle Erichsena.

9. Přehled a význam použitých symbolů a termínů

Symbol, termín	*Význam*
α	hladina významnosti
data, x	náhodný výběr
$f(x)$	hustota pravděpodobnosti
$F(x)$	distribuční funkce
$F^{-1}(x)$	kvantilová (inverzní distribuční) funkce
$\Gamma(z)$	gamma funkce
$\mathrm{h}(x)$	skoková funkce, $\mathrm{h}(x) = 0$ pro $x<0$, 1 jinak; $x \in \mathbb{R}^1$
$L(\theta)$	věrohodnost
μ	střední hodnota
$N(\mu,\sigma^2)$	normální rozdělení
$\mathbb{R}^n$	n-rozměrný reálný prostor
$S(\mathbf{\theta})$	Kritérium pro optimalizaci regresního modelu
σ	směrodatná odchylka
$\sigma(x)$	aktivační funkce neuronu
$\mathbf{\theta}$, $\boldsymbol{\alpha}$	vektor parametrů statistického modelu
$x_{(i)}$	i-tý prvek vzestupně setříděného náhodného výběru
x_i	i-tý prvek náhodného výběru
$\overline{x}$	aritmetický průměr
$\tilde{x}$, $\tilde{x}_{0.5}$	výběrový medián
$\tilde{x}_{\alpha}$	výběrový α-kvantil

10. Seznam příloh

Příloha 1: Popis a definice jazyka DARWin